The Influence of Statistical Context on the Neural Representation of Sound

Ross S. Williamson

Dissertation submitted for the degree of
Doctor of Philosophy
of
University College London

Gatsby Computational Neuroscience Unit
and
Centre for Mathematics and Physics in the Life Sciences and Experimental Biology

University College London

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Ross Stewart Williamson
February 21, 2012

Abstract

Models of stimulus-response functions have been used for decades in an attempt to understand the complex relationship between a sensory stimulus and the neural response that it elicits. A popular model for characterising auditory function is the spectrotemporal receptive field (STRF), originally due to Aertsen and Johannesma (1980); Aertsen et al. (1980, 1981). However, the STRF model predicts auditory cortical responses to complex sounds very poorly, presumably because the model is linear in the stimulus spectrogram and thus incapable of capturing spectrotemporal nonlinearities in auditory responses.

Ahrens et al. (2008a) introduced a multilinear framework, which captures neuron-specific nonlinear effects of stimulus context on spiking responses to complex sounds. In such a framework, contextual effects are interpreted as nonlinear stimulus interactions that modulate the input to a subsequent STRF-like linear filter. We derive various extensions to this framework, and demonstrate that the nonlinear effects of stimulus context are largely inseparable, and fundamentally different for near-simultaneous and delayed non-simultaneous sound energy. In two populations of neurons, recorded from the mouse auditory cortex and thalamus, we show that simultaneous sound energy provides a nonlinear positive (amplifying) gain to the subsequent linear filter, while non-simultaneous sound energy provides a negative (dampening) gain. We demonstrate that this structure is largely responsible for providing a significant increase in the predicitve capabilities of the model.

Using this framework, we show that nonlinear context dependence differs between cortical fields, consistent with previous studies (Linden et al., 2003). Furthermore, we illustrate how such a model can be used to probe the nonlinear mechanisms that underly the ability of the auditory system to operate in diverse acoustic environements. These results provide a novel extension to the study of receptive fields in multiple brain areas, and extend existing understanding of the way in which stimulus context drives complex auditory responses.

CONTENTS

LIST OF FIGURES

PREFACE

ACKNOWLEDGEMENTS

I feel privileged to have been supervised by two spectacular academics throughout the course of my graduate studies. In addition to being involved in every aspect of my doctoral work, Maneesh Sahani and Jennifer Linden have both been instrumental in shaping my scientific thinking and development as a young scientist. I am completely indebted to them for the support and encouragement that they have provided me with over the years. Maneesh is clearly capable of thinking in a far higher dimensional space than I am, and he has always been able to astound and amaze me with his rigorous technical knowledge, and deep scientific intuition. Jennifer was kind enough to provide me with experimental training in her lab at the UCL Ear Institute. For this, I am eternally grateful. Such an opportunity changed the course of my PhD, and allowed me to engage in true interdisciplinary science. On a personal level, both Maneesh and Jennifer have always been there whenever I've had a problem, academic or otherwise. I could not have wished for a better pair of supervisors, and I hope we can continue to collaborate in the future.

The Gatsby Computational Neuroscience Unit has been a very special home for the last four years, and I will miss it dearly. I have made a number of great friends since I've been here, and a lot of the enjoyment of my PhD has been due to them. The quality of the Unit owes much to its director, Peter Dayan, a consummate researcher, who ensures that Gatsby remains a stimulating research environment.

In Office 505, Phillipp Hehrmann and David Barrett have become close friends. They have been present for all my failures, and all my little successes. They have always been there to help out whenever I've had a problem. I hope we can all stay in touch.

Misha Ahrens has been a great collaborator, and a great friend. I'm incredibly grateful

that he was happy for me to carry on his torch, and continue to work on the multilinear framework that he developed during his time at the Gatsby Unit. Were it not for his expert advice and guidance in times of crisis, this thesis would not be what it is today.

Bjorn Christianson and Lucy Anderson have become wonderful friends over the years. They were both instrumental in providing me with training as an experimentalist, and teaching me much of what I now know of the auditory system. Lucy introduced me to the wonders of the thalamus, and made me realise how much better it is than cortex.

Jan Gasthaus has not only become a great friend, but a great cycling partner. I'll fondly remember the number of times that we near killed ourselves sprinting over the English countryside. Our trips to Majorca were stuff of legend.

I have met so many wonderful people during my time at UCL. Loic, Charles, Vinayak, Lars, Andriy, Biljana, Jannis, Roland: it's been an absolute pleasure. Thanks also go to the rest of my Gatsby/Ear Institute friends.

Finally, I would like to thank my parents. Throughout my every endeavour, they have never ceased to provide me with anything but their constant love and support. For this, and everything, I thank you. I hope this thesis does you proud.

Collaborations and Contributions

Maneesh and Jennifer have been involved with every aspect of this thesis. Without them, it would not exist.

Electrophysiological Data. The mouse cortical data used in chapter 4 was collected by Jennifer whilst she was at the University of California, San Francisco. All thalamic data in chapter 4, and all cortical and thalamic data in chapter 5, was collected by myself at the UCL Ear Institute. I am indebted to both Lucy Anderson and Bjorn Christianson for providing electrophysiological and histological assistance during these experimental sessions.

Theory. In chapter 3, material up to section 3.3.1 is a review of previous work. The remainder of the chapter (with the exception of the discussion of the ASD algorithm, which is due to Sahani and Linden (2003a)) was carried out in collaboration with Misha Ahrens. This includes the development of the extended context model, and the derivation and implementation of the variational approximation to bilinear systems.

Modelling. The observation of the contextual structure present within the multilinear framework was originally observed in cortex, and is due to Misha Ahrens (and briefly discussed in Ahrens et al. (2008a)). I helped to develop the ideas further and applied the framework to data recorded from the thalamus. All of the data analysis presented in chapter 4 was carried out by myself. The work in chapter 5 is my largely my own.

Other work during the PhD

Over my four years at the Gatbsy Unit, I have worked on a number of different projects all related the overarching theme of neural encoding. In order to provide a cohesive thesis with one specific theme, some of my PhD research has been omitted. The thesis itself will largely focus on work carried out during my final eighteen months in the Unit.

Experience Dependent Plasticity in Rat Auditory Cortex

Prior to starting my PhD proper, I gained an MRes in which my thesis project was carried out with Maneesh and Jennifer. I investigated the role of experience dependent plasticity in the cortical responses of the rat, using both linear and multilinear methods. This project was extended, and became the research focus of the first year of my PhD. It resulted in the following four conference proceedings:

> J.F. Linden, I. Orduna, **R.S. Williamson**, M.B. Ahrens, E. Mercado, M.M. Merzenich, M. Sahani (2009). Experience dependent shaping of complex response properties in adult auditory cortex. *British Society for Audiology, Short Papers Meeting*.
>
> J.F. Linden, I. Orduna, **R.S. Williamson**, M.B. Ahrens, E. Mercado, M.M. Merzenich, M. Sahani (2009). Experience dependent shaping of complex response properties in adult auditory cortex. *Auditory Cortex Meeting*.
>
> J.F. Linden, I. Orduna, **R.S. Williamson**, M.B. Ahrens, E. Mercado, M.M. Merzenich, M. Sahani (2009). Auditory learning involving complex sounds affects nonlinear integration within cortical responses. *Computational and Systems Neuroscience*.
>
> J.F. Linden, I. Orduna, **R.S. Williamson**, M.B. Ahrens, E. Mercado, M.M. Merzenich, M. Sahani (2009). Auditory learning involving complex sounds affects nonlinear integration within cortical responses. *Association for Research in Otolaryngology*.

Relating information-theoretic and likelihood-based methods for spike-triggered neural characterisation

I have also engaged in a more theoretical project in collaboration with Maneesh Sahani and Jonathan Pillow (at the University of Texas at Austin). This project focussed on understanding the mathematical links between two popular neural encoding methods; that of *maximally informative dimensions* and *linear nonlinear Poisson cascades*. This work resulted in the following conference proceeding, and is currently being prepared for journal submission:

> **R.S. Williamson**, M. Sahani, J.W. Pillow (2011).
> On information theoretic and likelihood based methods for spike-triggered neural characterisation. *Computational and Systems Neuroscience*.

Glossary of Acronyms

AAF	Anterior Auditory Field
A1	Primary Auditory Cortex
AII	Secondary Auditory Cortex
ALS	Alternating Least Squares
AN	Auditory Nerve
ARD	Automatic Relevance Determination
ASD	Automatic Smoothness Determination
BF	Best Frequency
CF	Characteristic Frequency
CGF	Contextual Gain Field
CNC	Cochlear Nucleus Complex
DCN	Dorsal Cochlear Nucleus
DRC	Dynamic Random Chord
GLM	Generalised Linear Model
IC	Inferior Colliculus
LNP	Linear-Nonlinear-Poisson
LSO	Lateral Superior Olive
MAP	Maximum A-Posteriori
MGB	Medial Geniculate Body
dMGB	Dorsal Division of the Medial Geniculate Body
mMGB	Medial Division of the Medial Geniculate Body
vMGB	Ventral Division of the Medial Geniculate Body
ML	Maximum Likelihood
MNTB	Medial Nucleus of the Trapezoid Body
MSO	Medial Superior Olive
NLL	Nucleus of the Lateral Lemniscus
NRC	Normalised Reverse Correlation
PRF	Principal Receptive Field
PSTH	Peri Stimulus Time Histogram
SOC	Superior Olivary Complex
STA	Spike Triggered Average
STC	Spike Triggered Covariance

STRF	Spectrotemporal Receptive Field
SVD	Singular Value Decomposition
UF	Ultrasonic Field
VCN	Ventral Cochlear Nucleus

For my parents,
for not making me get a real job.

I'll be forever grateful.

I

Introduction

1.1 Neural Encoding in Sensory Neuroscience

One of the fundamental goals within sensory neuroscience is the ability to successfully characterise the relationship between a sensory input, and its corresponding neural representation. This is known as neural encoding.

Over many decades much effort has been devoted to establishing the best neural encoding model to use in order glean insight into the elusive relationship between stimulus and response. Arguably, the simplest possible model to use for this purpose is a linear one and, as a result, they have been widely used throughout multiple sensory systems (e.g., Eggermont et al. (1983); Wu et al. (2006)). In the auditory domain, such models are known as spectrotemporal receptive fields (STRFs) and are linear in the spectrogram of the stimulus.[1] These have been applied to various stages in the auditory pathway, from the cochlear nucleus (Nelken et al., 1997; Steinberg and Peña, 2011), to the inferior colliculus (Escabi and Schreiner, 2002), thalamus (Miller et al., 2002), and auditory cortex (Depireux et al., 2001; Linden et al., 2003). However, other studies have explicitly looked to quantify just how much variability within the neural response such models are actually able to capture (Sahani and Linden, 2003b; Machens et al., 2004), with the result being, in cortex at least, that a linear model can typically account for no more than 20-40% of the stimulus-dependent variability in neural responses to complex

[1] The original development of STRF models is due to Aertsen et al. (1980); Aertsen and Johannesma (1980); Aertsen et al. (1981). Prior to this, de Boer and de Jongh (1978) introduced models which were linear in the sound pressure waveform, as opposed to the stimulus spectrogram which is more common these days.

sounds.

This failing of the linear model is largely due to the fact that the true relationship between the sensory stimulus, and its corresponding neural response is a highly nonlinear one. In the auditory cortex for example, neural responses have been shown to be strongly and nonlinearly modulated by stimulus context (Brosch et al., 1999; Brosch and Schreiner, 2000; Bartlett and Wang, 2005; Calford and Semple, 1995; Sadagopan and Wang, 2009; Bar-Yosef et al., 2002; Bar-Yosef and Nelken, 2007). One way to potentially deal with such nonlinearities is to increase the complexity of the model by, say, allowing second-order interactions to be captured, through the use of a Volterra series expansion (Marmarelis and Marmarelis, 1978). However, this flavour of nonlinear model suffers from the curse of dimensionality, as the amount of data needed to fit the model rises exponentially with the model order.

One way to approach this issue with dimensionality is to reduce the number of model parameters by tailoring a model to the observed properties of auditory neurons. Ahrens et al. (2008a) introduced multilinear "context" models, which capture neuron-specific nonlinear effects of stimulus context on spiking responses to complex sounds. In such a framework, contextual effects are interpreted as non-linear stimulus interactions that modulate the input to a subsequent STRF-like linear filter.

We show that, with various extensions to this framework, we can successfully estimate nonlinear interactions from the neural responses to complex sounds in both the auditory cortex and thalamus. This provides a novel extension to the study of receptive fields in multiple brain areas, and extends existing understanding of the way in which nonlinear stimulus context drives complex auditory responses.

1.2 Outline of the Thesis

This thesis proceeds as follows:

Chapter 2 provides a general introduction to neural encoding in the mammalian auditory system. We first briefly discuss some general auditory physiology and anatomy, before moving on to a discussion of spike-triggered neural characterisation, and the use of stimulus-response functions in auditory neuroscience.

Chapter 3 introduces the multilinear framework for modelling neural responses to sound. This chapter contains a combination of both background material and original work. We first summarise the earlier work of Ahrens et al. (2008a,b), who were responsible for introducing this framework. We then proceed to develop the framework further, by presenting an extended model that captures inseparable contextual effects. We also provide details about how to perform parameter estimation in such a model. This work was carried out in collaboration with Misha B. Ahrens, Maneesh Sahani, and Jennifer F. Linden.

Chapter 4 is the first of two primary results chapters within the thesis. We apply the extended context model to data from the mouse auditory cortex and thalamus, and show that we can successfully estimate nonlinear inseparable contextual interactions, that contain biological relevance. The predictive capabilities of such a model are also higher than what has been described previously. This work was carried out in collaboration with Misha B. Ahrens, Maneesh Sahani, and Jennifer F. Linden, and is currently in preparation for journal submission.

Chapter 5 is the second of two primary results chapters within the thesis. We record extracellular responses from mouse auditory cortex and thalamus, using a spectrotemporally-rich stimulus that varies in spectrotemporal density. We quantify various aspects of the neural responses themselves, before showing that the use of the extended context model can shed light on the nonlinear interactions present within the stimulus. This yields insight into how contextual processing plays a role in complex acoustic environments. This work was carried out in collaboration with Lucy A. Anderson, Maneesh Sahani, and Jennifer F. Linden, and is currently in preparation for journal submission. This work has also appeared, in various incarnations, in the following conference proceedings:

> **R.S. Williamson**, L.A. Anderson, G.B. Christianson, M. Sahani, J.F. Linden (2011). Auditory thalamic neurons show nonlinear sensitivity to stimulus context. *Society for Neuroscience*.
>
> **R.S. Williamson**, L.A. Anderson, G.B. Christianson, M. Sahani, J.F. Linden (2011). Auditory thalamic neurons show nonlinear sensitivity to stimulus context. *Advances and Perspective in Auditory Neurophysiology*.
>
> **R.S. Williamson**, L.A. Anderson, G.B. Christianson, M. Sahani, J.F. Linden (2011). Stimulus density dependence in the auditory thalamus. *Computational and Systems Neuroscience*.
>
> **R.S. Williamson**, L.A. Anderson, G.B. Christianson, M. Sahani, J.F. Linden (2011). Stimulus density dependence in the auditory thalamus. *Association for Reasearch in Otolaryngology*.

Chapter 6 concludes the thesis.

II

Neural Encoding in the Mammalian Auditory System

Outline

This chapter provides an overview of material relevant to the thesis. It starts off by briefly discussing the anatomy and physiology of the mammalian auditory system, with a focus on the flow of auditory information from the periphery to the cortex. From there, the chapter moves into a more mathematical direction and presents some previous work on mathematical and computational approaches to neural encoding (with a focus on the auditory system), more of which will be discussed in the coming chapters.

2.1 Anatomy and Physiology of the Mammalian Auditory System

The sense of hearing provides us with information about the world around us. In order to make sense of this information, the mammalian auditory system has the ability to accurately encode incoming acoustic information as trains of action potentials (or spikes) that propagate through multiple brain areas before reaching the neocortex.

This thesis is largely concerned with two such brain areas. The first of these, the medial geniculate body (MGB), is an obilgatory thalamic relay station that is responsible for transmitting information onto cortex. The second of these areas, the auditory cortex, is one of the final stages in the early auditory pathway and is thought to be responsible for more complicated tasks such as the processing of information conveyed within complex sounds.

This background section will first provide a brief overview of peripheral and central auditory processing, before providing a more detailed description of the auditory thalamus and cortex, around which this thesis is centered.

2.1.1 Peripheral Auditory Processing

The auditory pathway is inherently complex. It includes a large number of different processing stages, occurring both sequentially and in parallel, that serve to both modify and augment neural firing patterns until they reach their final processing stations in the auditory cortex.

When something makes a sound, oscillations of pressure (or sound waves) are sent through the air. These oscillations are collected by the visible part of the ear (known as the pinna). They are then funneled along the auditory canal towards the tympanic membrane (more commonly known as the ear drum) to which are connected the smallest bones in the body; ossicles. These ossicles act as a lever system, transferring the low-pressure movements of the eardrum into higher-pressure movements of a second membrane (the oval window) which, in turn, causes a movement of fluid within the cochlea.

The cochlea itself takes on a spiralled shape within which lies a hollow tube with walls made of bone. Interestingly, the only functional significance of this spiralled shape is to keep the space occupied by the relatively long structure to a minimum. This hol-

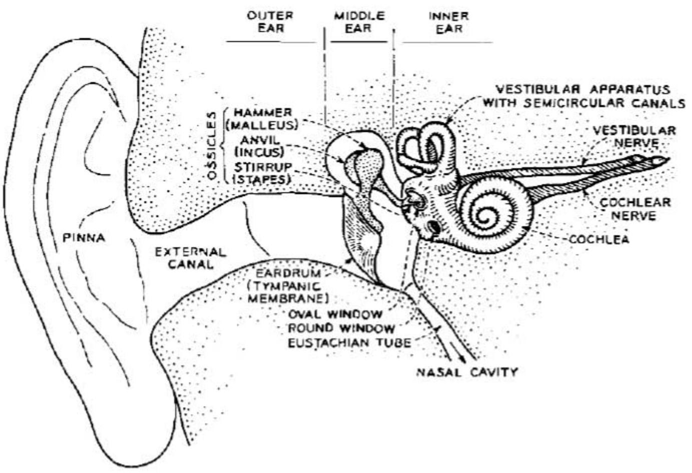


Figure 2.1: Peripheral auditory system (adapted from Flanagan (1972)). Illustration of the structure of the peripheral auditory system showing the outer, middle, and inner ear.

low tube contains three fluid filled chambers, known as scalae, which are separated by Reissner's membrane and the basilar membrane. The basilar membrane is a structure of great importance in the auditory system. It has two particularly important structural properties that mould its response to sound. Firstly, its dimensions vary along the length of the cochlea, becoming wider on moving from the base to the apex. Secondly, the basal end is incredibly rigid in contrast to the floppy apex. These structural properties have several repercussions for the way in which sound waves travel across the membrane (Von Békésy, 1980). In general, high frequency sound waves cause the rigid base to vibrate a great deal causing the sound energy to dissipate and, as a result, not travel very far along the membrane. In contrast to this, a low frequency sound wave will cause less vibration at the base and will allow the wave to travel further towards the apex.

Sitting atop the basilar membrane is the construct that contains the auditory receptors, the organ of Corti. These auditory receptors are known as hair cells, deriving their name from the bundle of projections (stereocilia) which protrude from their apical surface. There are two types of hair cells (inner and outer), which differ in size, shape, and function. The inner hair cells are the primary receptor cells, whose frequency selectivity is largely determined by their position on the basilar membrane. Deflection of these cells allows mechanically gated ion channels to be opened, resulting in a receptor potential due to the influx of positive ions (primarily potassium and calcium). In turn, this receptor potential opens voltage gated calcium channels, causing the release of neurotransmitters and the innervation of a set of spiral ganglion cells that surround the auditory nerve (AN). These spiral ganglion cells propagate action potentials along the AN and into the central auditory system.

2.1.2 Central Auditory Processing

2.1.2.1 The Brainstem

All AN fibers terminate at the cochlear nucleus complex (CNC), the first relay station of the ascending auditory pathway (Cant and Benson, 2003). The CNC also inherits the rich tonotopic structure generated by the systematic variation in mechanical properties of the basilar membrane, and the orderly arrangement of auditory nerve fibre dendrites along the length of this membrane (Arnesen and Osen, 1978). Such tonotopy is largely replicated throughout the auditory pathway (Clopton et al., 1974).

The CNC can be further subdivided into a ventral component; the ventral cochlear nucleus (VCN), and a dorsal component; the dorsal cochlear nucleus (DCN). The VCN is important for the temporal processing of sound. Several of its cell types are capable of transmitting precise temporal information, thus implicating the area in tasks such as sound localisation (Rhode et al., 1983). In addition, the VCN also contains so-called octopus cells, that are capable of encoding the pitch period of periodic sounds like vowels in their temporal firing patterns (Oertel, 1999). The majority of neurons in this area project to the superior olivary complex (SOC). The DCN is also involved in sound localisation in the vertical plane due to the apparent sensitivity of its neurons to spectral cues generated by the pinna (May, 2000). The area has also been implicated in the analysis of complex sounds (Young et al., 1992), and its neurons project to the inferior colliculus (IC) via the nuclei of the lateral lemniscus (NLL).

The CNC projects to the superior olivary complex (SOC), which is divided into three primary nuclei; the medial superior olive (MSO), the lateral superior olive (LSO), and the medial nucleus of the trapezoid body (MNTB). The MSO is the first area in the ascending auditory system that receives binaural input and, as such, MSO neurons are ideally suited for the measurement of interaural phase, or time, differences (Joris et al., 1998). By comparison, the LSO is primarily involved in interaural level difference detection, due to the input it receives from the MNTB (Tollin, 2003). These are both very important cues for sound localisation (McAlpine and Grothe, 2003; Grothe et al., 2010).

2.1.2.2 The Midbrain

The next stage in the ascending auditory pathway is the inferior colliculus (IC), located within the midbrain. The vast majority of ascending neural pathways synapse in this area (Aitkin and Phillips, 1984; Casseday et al., 2002). The IC itself is typically subdivided into a central core and several belt areas. The central core of the IC is a site of convergence for projections from more than twenty identified neurone types (Irvine, 1992). These different cell types have different functional properties, yet all terminate in a consistent, highly organised manner, providing the tonotopic structure within the central nucleus. In addition to tonotopy, other organisational arrangements have been suggested that include sound intensity, sound duration, frequency sweep direction, modulation rate, and other complex sound patterns (Langner and Schreiner, 1988; Schreiner and Langner, 1988). The IC is responsible for integrating information from the projections it receives, and therefore also has a role to play in sound localisation (e.g., Kuwada et al. (1979); Aitkin et al. (1985)). Strong adaptation to the distribution of sound level has also been observed (Dean et al., 2005).

Importantly, the IC also represents the principal source of information that ascends to the auditory thalamus. These projections are numerous, and have been established based on both functional properties and anatomical evidence (Morest, 1965; Andersen et al., 1980; Calford and Aitkin, 1983). Combined, these projections form the origin of parallel pathways which run through the auditory thalamus, and terminate in the auditory cortex.

2.1.2.3 The Thalamus

The principal nucleus of the auditory thalamus is known as the medial geniculate body (MGB), the other two areas being the lateral aspect of the posterior thalamic nucleus (PoL), and the auditory division of the reticular nucleus (Jones, 1985). The MGB itself can be futher subdivided into three divisions (ventral, medial, dorsal) on the basis of anatomy, histochemistry, and physiological responses, in a number of different species (e.g. mouse (Anderson et al., 2009a; Anderson and Linden, 2011), guinea pig (Anderson et al., 2007), cat (Calford, 1983), and monkey (Hackett et al., 1998)).

Classically, these three subdivisions have been attributed to either a primary (lemniscal) pathway, or a secondary (non-lemniscal) pathway, where the terms "lemniscal" and "non-lemniscal" are used to acknowledge whether or not the pathway includes the

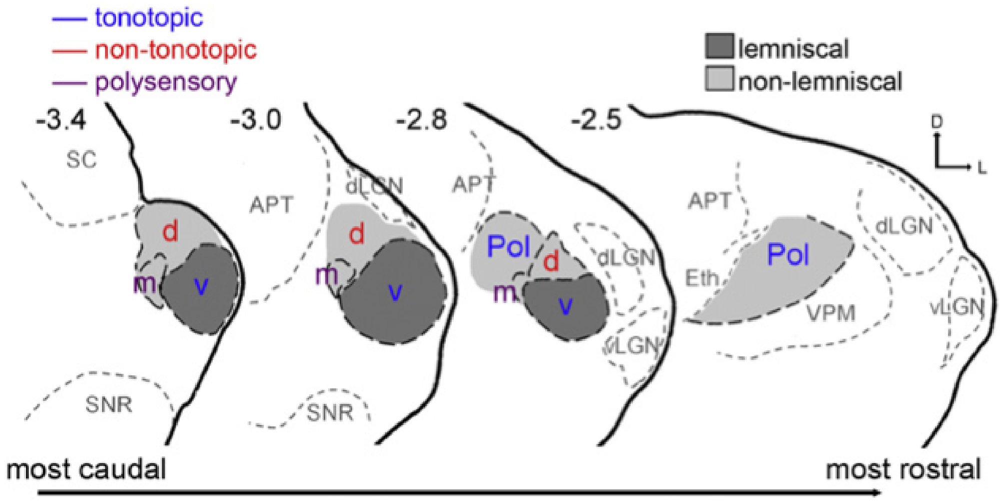


Figure 2.2: Thalamic anatomy (adapted from Anderson and Linden (2011). Line drawings of four coronal sections through a typical mouse thalamus to show the relative positions of the mouse MGB subdivisions and the POL. Borders were ascertained on the basis of histological delineation. Auditory areas are outlined in black dashed lines, and non-auditory areas in grey. Areas thought to belong to the lemniscal pathway are highlighted in dark grey, with areas areas thought to belong to the non-lemniscal pathway highlighted in light grey. Lettering colour also denotes pathway membership. Sections have a thickness of 40 μm, numbers at the top of each section indicate approximate distance (in mm) behind Bregma.

lateral lemniscus. The ventral division of the MGB (vMGB) is considered part of the primary pathway, and receives strong projections from the central nucleus of the inferior colliculus before projecting to layers III and IV of the auditory cortex (Winer et al., 2005). Both medial and dorsal subdivisions (mMGB and dMGB) are thought to be part of the secondary pathway, with both receiving input from all parts of the inferior colliculus (and other brain areas) before projecting to the non-primary (secondary) auditory cortices (Kimura et al., 2003; Winer et al., 2005). These different subdivisions and pathways are shown graphically in figure 2.2.

These multiple pathways have been traditionally thought to engage in different auditory functions. In fact, these pathways are sometimes referred to as being either “drivers”, or “modulators”, terms derived from pathways that carry (or *drive*) information, and pathways that *modulate* these principal information streams (Lee and Sherman, 2010).

Ultimately, it is the non-lemniscal pathways that prove the most interesting, from the perspective of understanding the processing of complex sounds. The lemniscal pathway provides a very “primary-like” representation of sound, in that tonotopy is

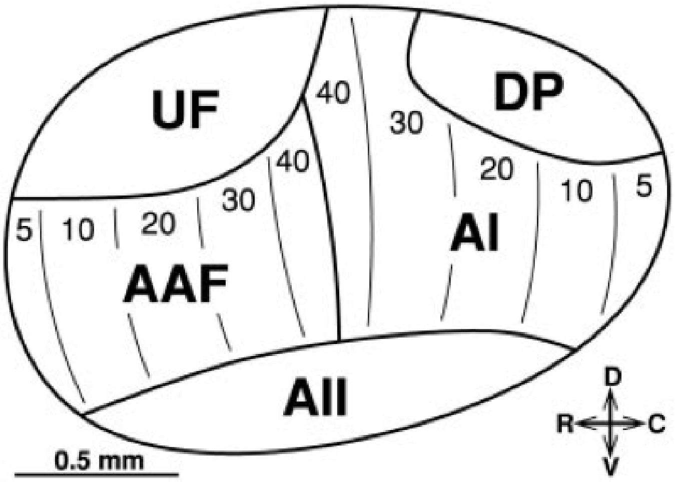


Figure 2.3: Cortical anatomy (adapted from Stiebler et al. (1997)). Line drawing of a typical mouse auditory cortex. A1, primary auditory cortex; AAF, anterior auditory field; UF, ultrasonic field; AII, secondary auditory field; DP; dorso-posterior field. Note the characteristic reversal in tonotopy along the rostral-caudal axis between A1 and AAF.

present throughout the pathway, and neurons respond well to classic auditory stimuli (clicks, modulated noise, frequency-modulated sweeps, etc) (de Ribaupierre, 1997). In contrast to this, the non-lemniscal pathways has been recently implicated in more context-dependent like responses, with neurons that show the ability to detect change (Ulanovsky et al., 2003, 2004; Anderson et al., 2009b; Malmierca et al., 2009).

2.1.2.4 The Auditory Cortex

The auditory cortex can be divided into a number of different subfields, that can be defined both tonotopically (Stiebler et al., 1997) and anatomically (Lee et al., 2004). Mice have five different cortical fields (Stiebler et al., 1997). Of these, two are considered to be *core*; the primary auditory cortex (A1) and anterior auditory field (AAF). Both of these core fields are tonotopically organised and share a high frequency reversal point at their separation boundary. A typical mouse auditory cortex is shown in figure 2.3.

Other species also have multiple cortical fields, including tonotopically organised core areas like A1 and AAF. For example, monkeys have a core region that includes three primary fields, surrounded by two different *belt* areas (de La Mothe et al., 2006). Rats are also similar to mice in so far as they have both an A1 and AAF, plus an additional four cortical fields (Polley et al., 2007).

The contribution of different cortical fields to auditory processing has been studied extensively in the literature. Previous studies have successfully mapped, across cortical fields, the neural sensitivity to a variety of different sounds, ranging from pure tones to complex structures varying in perceptual attributes such as pitch and timbre (Bizley et al., 2009; Bizley and Walker, 2009). It is also relatively well established that the contributions of such auditory fields are particularly important for perception. Behavioural relevance can be readily studied by employing inactivation techniques, either through permanent cortical lesions, or reversible inactivation using cooling methods (Lomber et al., 1999). Both techniques have demonstrated (for example) the importance of auditory cortex in sound localisation (e.g. Smith et al. (2004); Bizley et al. (2007); Lomber and Malhotra (2008)). More recently, the ability to reversibly inactivate cortical fields has provided insight into how cross-modal reorganisation of deaf auditory cortex can lead to enhanced visual performance (Lomber et al., 2010).

2.1.2.5 Processing of Complex Sounds

In some sense, the early stages in the pathway can be thought of as carrying out a form of cue detection, with rudimentary frequency and temporal coding being handled at the level of the AN and CN. The encoding of temporal information, in terms of phase locking at least, is maintained to some degree throughout all levels of the pathway, but degrades as it ascends. In a similar vein, ascending structures in the auditory system have increasingly poor frequency response specificity. As an example of this, the tuning curves of neurons at the level of the IC are typically far broader than those found in the CN, suggesting that an additional "recoding" of frequency information occurs during transmission. Of course, as soon as binaural convergence takes place, the situation becomes more complex still, due to the feature extracting properties of the SOC and its important role in sound localisation. The IC clearly has a complex and multiplexed function, with neurons that respond to a large variety of different auditory stimulus features. At the level of the auditory cortex itself, comparatively little is known. It is typically regarded as a "high" sensory area, with a vast array of complex functions.

It is my intent that chapters 4 and 5 of this thesis provide a significant step towards further understanding the auditory processing capabilities and functionality of both the auditory cortex and thalamus.

2.2 Spike-Triggered Neural Characterisation

2.2.1 Neural Dimensionality Reduction

There exists a large body of literature in computational neuroscience that focusses on methods for characterising the relationship between a sensory stimulus and the neural spike train that it elicits. Although this problem has been studied extensively in a variety of different sensory systems, the problem remains somewhat intractable due to the high-dimensional nature of the stimuli used in these kind of experiments. In an auditory experiment for example, even though an acoustic waveform is one-dimensional, it is then converted to some form of time-frequency representation (via an appropriate linear/nonlinear transform), which is inherently high-dimensional, containing sound energy at various points in both time and frequency. As a result of this, the vast majority of effort has focused on "neural dimensionality reduction" techniques as a means of estimating a low-dimensional subspace in which a neuron computes its response. The primary assumption underlying such an approach is that although the possible space of all stimuli is incredibly vast, the neural response will almost certainly not rely on all attributes of the stimuli. Thus, if a low-dimensional subspace in which a neuron computes its response can be identified, then the neural code can be characterised by describing responses only within that subspace.

2.2.2 Linear Stimulus-Response Functions

2.2.2.1 The Spectrotemporal Receptive Field

The simplest model that one can use to relate a sensory stimulus to a neural response is a linear one. In the auditory domain, this is known as a spectrotemporal receptive field (STRF) model, first described by Aertsen et al. (1980); Aertsen and Johannesma (1980); Aertsen et al. (1981). Such linear models have been widely used to characterise neurons within the auditory system (Nelken et al., 1997; Steinberg and Peña, 2011; Escabi and Schreiner, 2002; Miller et al., 2002; Depireux et al., 2001; Linden et al., 2003). Mathematically, an STRF model is given by

$$\hat{r}(i) = c + \sum_{jk} w_{jk}^{\mathbf{tf}} s(i - j + 1, k) \tag{2.1}$$

where $\hat{r}(i)$ is the firing rate at time i, $w^{\mathbf{tf}}$ is the STRF itself, and s is the stimulus. This equation describes a convolution in time (j is used to indicate time-lag indices) and a correlation in frequency (frequency being indexed by k) between the STRF and the time-frequency representation of the stimulus.

In the discrete-time framework, STRF estimation corresponds to a linear regression. The least-squares solution to such a regression problem is identical to the maximum likelihood (ML) value of $\mathbf{w}^{\mathbf{tf}}$ for a probabilistic regression model

$$r_i|\mathbf{x}_i \quad \sim \quad \mathcal{N}(\mathbf{w}^{\mathbf{tf}}\mathbf{x}_i, \sigma^2) \tag{2.2}$$

where $\mathbf{x}_i$ denotes a vector of stimulus intensities over some preceding time window that affects the spike response at time bin i. For a complete dataset with n stimulus-response pairs, the likelihood is given by

$$P(\mathbf{r}|X, \mathbf{w}^{\mathbf{tf}}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{r} - X\mathbf{w}^{\mathbf{tf}})^T(\mathbf{r} - X\mathbf{w}^{\mathbf{tf}})\right) \tag{2.3}$$

where $\mathbf{r}$ is a column vector of neural responses $\mathbf{r} = [r_1, r_2, \cdots, r_t]^T$, and X is a stimulus lag matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t]^T$, with the i^{th} row equal to $\mathbf{x}_i^T$.

The ML STRF estimate is then readily given by

$$\hat{\mathbf{w}^{\mathbf{tf}}} = \arg\max_{\mathbf{w}^{\mathbf{tf}}} P(\mathbf{r}|X, \mathbf{w}^{\mathbf{tf}}) = (X^TX)^{-1}X^T\mathbf{r} \tag{2.4}$$

where the RHS of equation 2.4 can be recognised as the ordinary least squares solution to a linear regression problem.

It is worth noting that, as in linear regression, this solution can also be regularised (this will be discussed at greater depth in chapter 3). This regularisation is typically carried out by adding a penalty term to the function being minimised; $(\mathbf{r} - \hat{\mathbf{r}})^T(\mathbf{r} - \hat{\mathbf{r}}) + {\mathbf{w}^{\mathbf{tf}}}^T\mathbf{D}\mathbf{w}^{\mathbf{tf}}$, where $\mathbf{D}$ is a matrix that contains coefficients for penalising terms in $\mathbf{w}^{\mathbf{tf}}$ (if $\mathbf{D} = \lambda\mathbf{I}$, this is simply ridge regression).

The STRF estimator given above is similar to the notion of a *spike-triggered average*, a common technique whereby the STRF of a neuron is estimated by averaging all of the stimulus components that precede spikes over the course of an experimental session. Mathematically,

$$\text{STA} \propto \sum_{t_{spk}} \mathbf{x}_{t_{spk}} = X^T\mathbf{r} \tag{2.5}$$

where $\mathbf{x}_{t_{spk}}$ is the stimulus segment preceding a spike at time t_{spk}. In practice, these times t_{spk} are binned. If there is more than one spike in a bin, then the stimulus vector for that time bin is multiplied by the number of spikes that occurred.

Thus, we can see that (under certain conditions), an STRF $\mathbf{w^{tf}}$ is identical to an STA in the case where $(X^T X) \propto I$, which is what happens if the stimulus is "white" (devoid of second-order correlations). This issue of correlations will be discussed in the next section.

2.2.2.2 Stimulus Dependence of STRFs

In recent decades STRFs have gradually attracted significant interest as a candidate framework for characterising auditory function (Kowalski et al., 1996a,b; deCharms et al., 1998; Linden et al., 2003). Despite their simplicity and interpretability however, they are not without problems. It has become well known that STRFs are stimulus dependent; that is, the nonlinearities present within neural response functions can lead to differences in STRFs estimated using different stimuli. Such problems have long been acknowledged in the literature (Marmarelis and Marmarelis, 1978; Aertsen and Johannesma, 1981; Theunissen et al., 2000; Escabi and Schreiner, 2002).

As a result of this, a lot of early work typically utilised white noise as a driving stimulus (Marmarelis and Marmarelis, 1978). This is a significant issue at higher levels in the auditory system however, where white noise tends to elicit very poor neural responses (Wang et al., 2005). This provoked the use of two primary types of stimuli, largely uncorrelated by design, in dynamic random chords (DRCs) (deCharms et al., 1998; Rutkowski et al., 2002; Linden et al., 2003), and temporally orthogonal ripple combinations (TORCs) (Kowalski et al., 1996a,b; Klein et al., 2000; Fritz et al., 2003). The uncorrelated nature of these stimulus designs typically allows for simple estimation of the STRF.

There has also been interest in STRF estimation with natural sounds (e.g., Theunissen et al. (2000); Sen et al. (2001); Woolley et al. (2005)). This can be challenging however, since natural stimuli typically contain a large amount of autocorrelation. Because of this, some dimensions have much lower variance than others (a natural sound will only probe a small region of the full stimulus space) (Theunissen et al., 2000, 2001). Such low variance dimensions provide very little modulating energy and thus make it very difficult to measure correlations between those dimensions and the response.

In equation 2.4, the general ordinary least squares solution for a linear STRF problem was shown. There, the inverse autocorralation matrix $(X^T X)^{-1}$ acts to normalise the variance along each dimension to be the same, therefore bringing the effect of the stimulus close to that of white noise. When variance is low however, normalisation requires division by a small number which can lead to an amplification of noise in parameter estimates. In short; inverting the stimulus autocorrelation matrix of natural sounds can be difficult and can lead to ill-conditioned matrices. Theunissen et al. (2001) proposed a solution to this, by using a pseudo-inverse to compute the stimulus autocorrelation inverse, and then setting dimensions below some noise threshold to zero. This has become known as the normalised reverse correlation (NRC) algorithm and has been used successfully to describe the receptive fields of auditory neurons in response to natural sounds (e.g., Woolley et al. (2005); Greene et al. (2009)).

First, an eigendecomposition is applied to the autocorrelation matrix

$$(X^T X) = C = U \Sigma U^T \tag{2.6}$$

where the columns of U contain the eigenvectors of C and the diagonal elements of Σ, (diag$(\lambda_1, \lambda_2, \cdots, \lambda_y)$), contain the corresponding eigenvalues, ordered by size. The number of dimensions to retain, less than some threshold τ (typically set by some cross-validation procedure) is given by

$$d = \arg\max \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_d}{\lambda_1 + \lambda_2 + \cdots + \lambda_d + \cdots + \lambda_y} < \tau \tag{2.7}$$

The pseudoinverse can then be calculated as

$$C^{-1}_{\text{approx}} = U \Sigma^{-1}_{\text{approx}} U^T = U \text{diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \cdots, \frac{1}{\lambda_y}, 0, \cdots, 0\right) U^T \tag{2.8}$$

and the final estimate of the STRF is given by

$$\mathbf{w^{tf}} = C^{-1}_{\text{approx}} X^T r \tag{2.9}$$

In addition to the NRC approach, variants based on the statistical technique of "boosting" (Friedman et al., 2000; Zhang and Yu, 2005) have also been used. These approaches were compared by David et al. (2007).

Unfortunately however, even though these techniques allow for efficient estimation of STRFs, they are still not able to alleviate the problem of stimulus dependence. The issue is that such an approach, whereby the correlational structure of the stimulus is whitened, is only able to eliminate second-order moments from within the stimulus. Christianson et al. (2008), in a series of simulations, illustrate that the presence of non-zero third, or higher, order correlations between elements in spectrotemporal space can produce elements within the STRF structure to which the neuron is actually insensitive.

This certainly does not discount STRF analysis as a valuable resource for studying auditory function. Even if there is correlational structure within the stimulus that cannot be discounted, provided the experimental design is adequate, this does not have to be a problem. An example of this is the work of Fritz et al. (2003, 2007); Elhilali et al. (2007), where they use TORCs to estimate STRFs under a selection of different behavioural conditions. Even though TORCs contain higher-order structure that will cause stimulus dependence, the stimuli are fixed for the different behaving and non-behaving conditions, so any difference between the linear STRF fits is still indicative of a real functional change.

2.2.2.3 A Geometric Perspective, and Nonlinear Extensions

This class of models is particularly elegant when considered from a geometric perspective, and is related to the notion of neural dimensionality reduction that was introduced earlier. Figure 2.4 (a) shows the geometric interpretation of the afore mentioned STA. Each of the blue dots represent a particular stimulus segment that was presented at some point during the experiment (obviously, the true stimulus segment will be high-dimensional, so only the two-dimensional representation is shown here for simplicity). Combined, the blue dots yield the empirical stimulus distribution $p(\text{stim})$. The red dots correspond to those stimulus segments that elicited a spike and yield the empirical conditional stimulus distribution $p(\text{stim}|\text{spike})$. The STA represents the average stimulus preceding a spike and thus, the STA denotes a direction within this space (indicated by the black line in the diagram) that corresponds to the difference in mean between the two distributions.

This kind of *moment-based estimation* is common in the literature as a way of identifying a low-dimensional subspace (in the above example; the STA) in which the neuron computes its response Chichilinsky (2001); de Ruyter van Steveninck and Bialek (1988); Schwartz et al. (2006). Of course, the mean is not the only moment that one can utilise. Another well-known technique, *spike-triggered covariance* (STC), works by maximising the difference in second moment. This technique was first conceived by de Ruyter van

Steveninck and Bialek (1988); Brenner et al. (2000), and has since been used extensively in many different sensory systems (see, for example Rust et al. (2004)).

We can define both the stimulus covariance and spike-triggered covariance matrix as follows. Defining $\mathbf{x}_{\text{sta}} = \frac{1}{N}\sum_{t_{spk}} x(t_{spk})$ and $\mathbf{x}_{\text{stim}} = \frac{1}{T}\sum_{t} x(t)$, where N is the total number of spikes and T is the number of time bins, the covariances are denoted [1]

$$C_{\text{spk}} = \frac{1}{N}\sum_{t_{spk}} \left(\mathbf{x}(t_{spk}) - \mathbf{x}_{\text{sta}}\right)\left(\mathbf{x}(t_{spk}) - \mathbf{x}_{\text{sta}}\right)^T \tag{2.10}$$

$$C_{\text{stim}} = \frac{1}{T}\sum_{t} \left(\mathbf{x}(t) - \mathbf{x}_{\text{stim}}\right)\left(\mathbf{x}(t) - \mathbf{x}_{\text{stim}}\right)^T \tag{2.11}$$

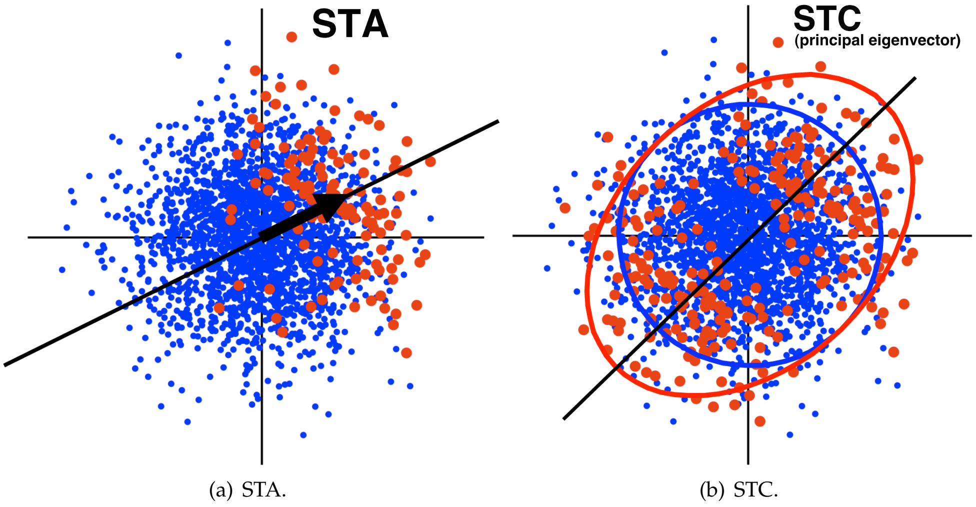
One is typically interested in identifying a set of directions whereby the variance of the spike-triggered stimuli differs maximally from that of the raw stimuli. Thus, a difference matrix based on the difference in second order structure can be defined as

$$C_{\text{diff}} = C_{\text{stim}}^{-\frac{1}{2}}\left(C_{\text{spk}} - C_{\text{stim}}\right)C_{\text{stim}}^{-\frac{1}{2}} \tag{2.12}$$

By performing an eigendecomposition on this difference matrix, the resultant eigenvectors can be used to define a low-dimensional subspace of interest. Here, eigenvectors associated with positive eigenvalues tend to correspond to stimulus features that make the neuron spike, while eigenvectors associated with negative eigenvalues correspond to stimulus features that suppress neuronal firing. This is shown graphically in figure 2.4 (b).

Another possibility is to use a divergence measure that is grounded in information theory. One such method seeks to find *maximally informative dimensions* (Sharpee et al., 2004; Paninski, 2003) whereby dimensions are found such that the information between stimulus and response is maximal (this amounts to maximising a Kullback-Liebler divergence between the two stimulus ensembles). This seems like a particularly attractive possibility since, in principle, such a technique is sensitive to statistical changes of any order (rather than just looking for differences in either mean or variance). Such a method has seen recent use within the auditory literature, in an attempt to identify nonlinearities over different cortical laminae (Atencio et al., 2008, 2009).

[1] Again, the t_{spk} here are typically binned in practice, and this means that each term should be multiplied by the number of spikes occurring in the associated time bin



(a) STA.

(b) STC.

Figure 2.4: STA/STC geometry. (a): A geometric representation of a spike triggered average. The blue dots each represent one of the many stimulus segments that were presented throughout the course of an experiment (projected into a two-dimensional space for visualisation). The red dots correspond to those stimulus segments that elicited a neural response. The STA yields the average of these red dots and thus represents a linear subspace within this space, illustrated by the black line. (b): Blue and red dots are as in (a). The elipses represent the covariance of each ensemble. The black line illustrates the direction in which the variance of each ensemble differs most.

2.2.3 Generalised Linear Models

In the linear models discussed earlier, minimising the squared error is an appropriate objective function to use. If one is dealing with an actual spike train however, as opposed to a PSTH, then a different objective function has to be used (it is easy to see that a Gaussian distribution is not a particularly good model of a binary variable).

Linear-nonlinear cascade models have become a popular way to approach spike-triggered neural characterisation in recent years. These models define the response in terms of a cascade of linear and nonlinear stages, followed by a probabilistic spiking process. The linear stage in the cascade is identical to what has been previously discussed, whereby a linear receptive field $\mathbf{w^{tf}}$ acts to reduce the dimensionality of the high-dimensional stimulus $\mathbf{S}$. After this dimensionality reduction step however, the aim is to find some function that relates the stimulus projection in this low-dimensional space to the actual probability of response. This is typically realised in the form of a static nonlinearity which, when applied to the filtered stimulus, is used to generate some form of probabilistic spiking.

The general form for a linear-nonlinear-Poisson model is given by

$$P(r|\lambda(X)) = \frac{\lambda(X)^{\mathbf{r}}}{\mathbf{r}!} \exp\left(-\lambda(X)\right) \tag{2.13}$$

where $\lambda(X) = f({\mathbf{w}^{\mathbf{tf}}}^T X)$, the intensity function of an inhomogeneous Poisson process. Here, the output of the cascade is Poisson, and thus the dimensionality reduction and static nonlinearity steps act as the intensity function for the Poisson process. One of the reasons for the popularity of this cascade framework is the existence of simple and computationally efficient fitting algorithms. The simplest of these involves using the STA as an estimate for $\mathbf{w}^{\mathbf{tf}}$, and a simple density estimation procedure for estimating the nonlinearity f (see Chichilinsky (2001) for details). Alternatively, one can express the likelihood of the model, which is relatively straightforward due to the fact that the PSTH bins are conditionally independent of one another given the stimulus, an essential feature of Poisson processes. Fixing the static nonlinearity to be some convex function results in the log-likelihood being concave, ensuring that the likelihood has no non-global local maxima (Paninski, 2004).

Another particularly appealing feature of such models, is that a variety of more general covariates can be included within the intensity function. In the standard LNP model, the intensity function takes the form

$$\lambda(t) = \lambda_{\text{stim}} = {\mathbf{w}^{\mathbf{tf}}}^T X \tag{2.14}$$

Typical extensions of this framework include the addition of terms to account for stimulus history and cross-neuron couplings

$$\begin{aligned} \lambda(t) &= \lambda_{\text{stim}} + \lambda_{\text{history}} + \lambda_{\text{coupling}} \\ &= {\mathbf{w}^{\mathbf{tf}}}^T X + \mathbf{h}^T \mathbf{y} + \sum_i \left(\mathbf{l}_i^T \mathbf{y}_i\right) + \mu \qquad (2.15) \end{aligned}$$

where $\mathbf{S}$ is the stimulus, $\mathbf{y}$ is the cell's own spike-train history, μ is a baseline firing rate, and i is used to index over other neurons with a population, such that $\{\mathbf{y}_i\}$ are the spike-train histories of the other cells. Importantly, even with this additional information, the resultant likelihood is still concave, and thus it is easy to find the maximum likelihood solution to the parameters. More complicated extensions to this framework have also been proposed; for example, models that contain stochastic, stimulus-dependent transitions (Escola et al., 2011) and models that contain more realistic single neuron dynamics (Paninski et al., 2004).

These kind of generalised linear models are particularly powerful and, as a result of their flexibility, they have been used extensively in the literature for a multitude of different tasks. These include studying the effect of correlations in retinal ganglion cells (Pillow et al., 2008), online decoding for motor prosthetics (Shoham et al., 2005), population dynamics and theta rhythms in the hippocampus (Harris et al., 2003), analysing functional connectivity (Okatan et al., 2005), and many more. Such a model has also seen recent attention in the auditory literature where it has been used as an alternative to the NRC algorithm discussed earlier (Calabrese et al., 2011).

III

Mathematical Aspects of Modelling Neural Responses to Sound

Outline

This chapter consists of a combination of both background material and original work. We start by summarising the work of Ahrens et al. (2008a,b) who were the first to introduce a multilinear framework for modelling neural responses. The material presented up to section 3.3.1 follows that of Ahrens et al. (2008a). In section 3.3.2, we proceed to develop this framework further, by presenting extended variants along with details of how to perform parameter estimation. This work was carried out in collaboration with Misha B. Ahrens (as well as Maneesh Sahani and Jennifer F. Linden). This chapter will largely treat the mathematical aspects of multilinear modelling. The biological relevance of the models, and applications to neural data will be provided in the later chapters of the thesis.

3.1 INTRODUCTION

The notion of neural encoding (at least from a functional perspective) is about understanding the complex relationship between a neural response and the sensory stimulus that drives it. From a theoretical standpoint, this amounts to modelling a "stimulus-response function"; the functional mapping from a sensory stimulus S to a vector of instantaneous firing rates $\mathbf{r}$. The simplest model of neural encoding is a linear one. As discussed in chapter 2, this is known in the auditory domain as a spectrotemporal receptive field (STRF). Such a model is computationally very simple, and easy to estimate, but it does have problems. The main problem, inherent within the model definition, is that it is linear. As a result of this, such a model is incapable of including effects such as history dependence or interactions amongst neurons. The linear model tells us that the contribution of a particular frequency is always the same irrespective of whether there were preceding or simultaneously presented frequencies. That is, given a combination of tones, a linear model states that the response to the combination can be predicted from the responses to the individual constituents of the combination. This is not necessarily the case. Here we present a framework of neural encoding models that are capable of capturing such nonlinear effects, present within neural responses throughout the auditory system. Formally, this framework is similar to a *Hammerstein* cascade (Hunter and Korenberg, 1986; Narendra and Gallman, 1966); however, its development in the multilinear setting, and its application to neural data, is more recent (Ahrens et al., 2008a,b).

3.1.1 NOTATION

For consistency, we will utilise the same notation as introduced by Ahrens et al. (2008a) (and also Englitz et al. (2010)). The models that will be discussed throughout this chapter are typically described through the use of multi-dimensional arrays (denoted by bold-faced letters such as $\mathbf{Q}$). Bold superscripts are used to specify the physical dimensions of such an array, and italicised subscripts to denote specific elements within the array. As it is auditory data that will be primarily dealt with, these physical dimensions typically correspond to time and/or frequency (although the modelling ideas can be easily extended to other sensory systems). Such a time-frequency array (essentially, an STRF) can thus be denoted as $\mathbf{w}^{\mathbf{tf}}$, (or separably as a time vector $\mathbf{w}^{\mathbf{t}}$ or a frequency vector $\mathbf{w}^{\mathbf{f}}$). Component notation can also be utilised to specify a single entry from within one of these arrays; $w^{\mathbf{tf}}_{jk}$ is the jk^{th} element in the array $\mathbf{w^{tf}}$.

Using this notation, the symbol $\otimes$ is used to generalise the vector outer product, such that, for vectors $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$

$$\mathbf{a} = \mathbf{b} \otimes \mathbf{c} \otimes \mathbf{d} \tag{3.1}$$

is a three-dimensional array, with the elements

$$a_{ijk} = b_i c_j d_k \tag{3.2}$$

In a similar vein, the symbol $\bullet$ is used to generalise the vector inner product, so that indices shared on the left and right hand sides of the operator are contracted. As an example, if we have a three-dimensional array a_{ijk}, and a two-dimensional array b_{jk}, then

$$\mathbf{c} = \mathbf{a} \bullet \mathbf{b} \tag{3.3}$$

is a vector (one-dimensional array) with

$$c_i = \sum_{jk} a_{ijk} b_{jk} \tag{3.4}$$

3.2 An Introduction to Multilinear Models

3.2.1 Bilinear Spectrotemporal Receptive Fields

As discussed in chapter 2, the simplest model that can be used to describe the functional relationship between stimulus and response is a linear one. In the auditory literature, this is typically known as the spectrotemporal receptive field (STRF), and is based upon the notion of reverse correlation. Such a model can be denoted as

$$\hat{r}(i) = \sum_{j=1}^{J} \sum_{k=1}^{K} w^{\mathbf{tf}}_{jk} s(i - j + 1, k) \tag{3.5}$$

where $\hat{r}(i)$ is the estimated instantaneous firing rate at time i, $w^{\mathbf{tf}}_{jk}$ is the jk^{th} element of the time-frequency STRF, and $s(i - j + 1, k)$ denotes an element from the stimulus spectrogram s at time-lag $j + 1$, and frequency k. The summation limits run from 1 to some maximal value of time-lag J and frequency K. For simplicity, we will suppress summation limits throughout the rest of the chapter. Unless stated otherwise, all summations run from 1 to some maximum value of that particular variable.

Equation 3.5 assumes an inseparable (or full-rank) representation of the receptive field $\mathbf{w^{tf}}$. Assuming separability in both time and frequency, an alternative model can also be written as

$$\hat{r}(i) = \sum_{jk} w^{\mathbf{t}}_j w^{\mathbf{f}}_k s(i - j + 1, k) \tag{3.6}$$

where the (now separable) receptive field $\mathbf{w^{tf}}$ is written as the outer product of a time vector and a frequency vector, $\mathbf{w^{tf}} = \mathbf{w^t} \otimes \mathbf{w^f}$. From a functional perspective, this notion of separability is useful since it implies that the response of the neuron to tones of different frequencies is preserved across time. Separability has been investigated frequently within the literature (Depireux et al., 2001; Klein et al., 2006), and is particularly useful statistically since it is associated with a substantial reduction in free parameters (from $(J \times K)$ to $(J + K)$). In addition, it is also a relatively biologically plausible approximation to make, since many core auditory cortical cells can be adequately classified as separable (Linden et al., 2003).

The model described by equation 3.6 is *bilinear*; it is linear in either of the given parameter vectors. It is this model which will provide the basic framework for discussing the multilinear models to come.

Here, we will make a slight simplification to notation. We define an expanded three-dimensional stimulus array, augmented by the addition of a time-lag dimension

$$M^{\mathbf{itf}}_{ijk} = \begin{cases} s(i - j + 1, k), & i \leq I, j \leq J, i - j \geq 0, k \leq K \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

which allows us to rewrite the bilinear system of equation 3.6 as

$$\hat{r}(i) = \sum_{jk} w^{\mathbf{t}}_j w^{\mathbf{f}}_k M^{\mathbf{itf}}_{ijk} \tag{3.8}$$

Using the generalised outer and inner products that were defined in section 3.1.1 the bilinear system can be written in the simplified form

$$\hat{\mathbf{r}} = (\mathbf{w^t} \otimes \mathbf{w^f}) \bullet \mathbf{M^{itf}} \tag{3.9}$$

For completeness, we can also define a simple extension of the linear model by dealing with the case of a neuron being spontaneously active in the absence of a sensory stimulus. We can thus incorporate a constant offset, c, which can be thought of as such a spontaneous rate. The model above becomes

$$\hat{\mathbf{r}} = c + (\mathbf{w^t} \otimes \mathbf{w^f}) \bullet \mathbf{M^{itf}} \tag{3.10}$$

This constant offset c can be incorporated cleanly into the bilinear framework via an appropriate augmentation of the stimulus array. Thus, we consider a new array $\mathbf{Q^{itf}}$ such that

$$Q^{\mathbf{itf}}_{ijk} = \begin{cases} M^{\mathbf{itf}}_{ijk} & i \leq I, j \leq J, i-j \geq 0, k \leq K \\ 1 & i \leq I, j = J+1, k = K+1 \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

where the stimulus array has been extended by one additional dimension.

This allows us to remove the explicit spontaneous rate term from equation 3.10 and gives us our final bilinear model

$$\hat{\mathbf{r}} = (\mathbf{w^t} \otimes \mathbf{w^f}) \bullet \mathbf{Q^{itf}} \tag{3.12}$$

Augmenting $\mathbf{w^t}$ and $\mathbf{w^f}$ to contain $J+1$ and $K+1$ elements respectively, the models defined by equations 3.10 and 3.12 become equivalent with $c = w^{\mathbf{t}}_{J+1} w^{\mathbf{f}}_{K+1}$.

3.2.2 Multilinear Models for Capturing Input Nonlinearities

Section 3.2.1 showed that a separable STRF can be cast in a multilinear framework. Such a bilinear model can be seen as the simplest, non-trivial multilinear model. As we will see as this chapter progresses, this multilinear framework can become far more complex, and capable of capturing realistic nonlinearities.

The general form for a multilinear model can be written as

$$\hat{\mathbf{r}} = (\mathbf{a} \otimes \mathbf{b} \otimes \cdots \otimes \mathbf{z}) \bullet \mathbf{Q} \tag{3.13}$$

where $\mathbf{a}, \mathbf{b}, \cdots, \mathbf{z}$ are arbitrary vectors of free parameters and $\mathbf{Q}$ is a fixed multidimensional array.

The time-frequency representation of a sound requires some form of linear/nonlinear operation on the sound pressure waveform (Gill et al., 2006). Thus, the scaling of the stimulus representation can severely influence the match between a given model and the data. Rather than assuming some fixed scaling, Ahrens et al. (2008b) use the multilinear framework to define an *input nonlinearity* model which aims to infer such a nonlinear transform directly from the data.

Building on the separable STRF model of equation 3.6, such a model takes the form

$$\hat{r}(i) = c + \sum_{jk} w^{\mathbf{t}}_j w^{\mathbf{f}}_k g(s(i-j+1,k)) \tag{3.14}$$

Here, the mapping g is this *input nonlinearity*, which acts to transform the representation of a sound level in the spectrogram prior to it being spectro-temporally filtered by the STRF. To allow for estimation, the mapping g has to be parametrised. A suitable choice is simply a linear combination of a fixed set of basis functions $\{g_l\}$, so that $g(x) = \sum_l w^{\mathbf{l}}_l g_l(x)$, for some parameter vector $\mathbf{w}^{\mathbf{l}}$. This yields

$$\hat{r}(i) = c + \sum_{jkl} w^{\mathbf{t}}_j w^{\mathbf{f}}_k w^{\mathbf{l}}_l g_l(s(i-j+1,k)) \tag{3.15}$$

The use of such a representation essentially reduces the problem of inferring this nonlinear stimulus transform to estimating the coefficients $\mathbf{w}^{\mathbf{l}}$ of the basis function set $\{g_l\}$.

As before, this model can be written in multilinear form. If we define a four-dimensional stimulus array $M^{\mathbf{itfl}}_{ijkl} = g_l(s(i-j+1,k))$, we can then write

$$\hat{r}(i) = c + \sum_{jkl} w^{\mathbf{t}}_j w^{\mathbf{f}}_k w^{\mathbf{l}}_l M^{\mathbf{itfl}}_{ijkl} \quad \text{or} \quad \hat{\mathbf{r}} = (\mathbf{w}^{\mathbf{t}} \otimes \mathbf{w}^{\mathbf{f}} \otimes \mathbf{w}^{\mathbf{l}}) \bullet \mathbf{Q}^{\mathbf{itfl}} \tag{3.16}$$

with a four-dimensional array $\mathbf{Q}^{\mathbf{itfl}}$ defined by augmenting $\mathbf{M}^{\mathbf{itfl}}$ in a manner analogous to equation 3.11.

3.3 Multilinear Models for Capturing Acoustic Context

3.3.1 The Fully-Separated Context Model

The input nonlinearity model defined in equation 3.15 describes a way in which the multilinear framework can be used to extend a simple linear neural encoding model, such that it is capable of incorporating an arbitrary nonlinear transformation of the sensory input. However, as attractive as this is, in both the STRF and the input nonlinearity model, stimulus features at different times or frequency are only ever combined linearly. That is, multiple features are never combined in anything other than a weighted, additive fashion. The vast majority of neural response nonlinearities are non-additive, and thus the biological plausibility of this model needs to be examined in more detail.

To address this issue Ahrens et al. (2008a) present an extension to this input nonlinearity model, wherein the multilinear framework is utilised to capture nonlinear "contextual" interactions. This extension is known as the context model, and will form the basis for the work presented within this thesis.

In its essence, the context model extends the previous models by additionally allowing a limited set of second order interactions. Intuitively, these interactions can be thought of as a short-term acoustic context; a contextual neighborhood surrounding each tone pulse within the stimulus. A contextual value is computed by weighting the tone pulses within such a small neighborhood, and then multiplying each value in the stimulus spectrogram by this sum. Finally, an STRF-like array is applied in order to produce the response in the form of an instantaneous firing rate. Specifically, the stimulus at time i with frequency k is given a strength denoted by

$$g(s(i,k)) \bullet (c_2 + \text{Context}(i,k)) \tag{3.17}$$

where the first term $g(s(i,k))$ is that of the input nonlinearity model discussed earlier, that will infer the effective level of a given tone pulse within the stimulus. This inferred level is then multiplicatively modulated by a context term given by

$$\text{Context}(i,k) = \sum_{\substack{mnp \\ (m,n)\neq(1,\Phi)}} w_m^{\boldsymbol{\tau}} w_n^{\boldsymbol{\phi}} w_p^{\boldsymbol{\lambda}} h_p(s(i-m+1, k-\Phi-1+n)) \tag{3.18}$$

where $\Phi = (N-1)/2$ dictates the maximum difference in frequency between the contextual and modulated time-frequency elements. The condition that $(m,n) \neq (1,\Phi)$ is to ensure that a tone does not appear within its own context.

Again, this contextual modulation can be succinctly expressed in multilinear notation. We can define a *contextual subunit*

$$[M^{\boldsymbol{\tau\phi\lambda}}(i,k)]_{mnp} = \begin{cases} 0 & \text{if} \quad (m,n) = (1,\Phi) \\ M^{\mathbf{itfl}}_{im(k-\Phi-1+n)p} & (otherwise) \end{cases} \tag{3.19}$$

where $M^{\boldsymbol{\tau\phi\lambda}}(i,k)$ is a stimulus array which depends on the ik^{th} position of the time-frequency element being modulated. Using this stimulus representation, we can now denote the contextual modulation of equation 3.18 as

$$\text{Context}(i,k) = (\mathbf{w}^{\boldsymbol{\tau}} \otimes \mathbf{w}^{\boldsymbol{\phi}} \otimes \mathbf{w}^{\boldsymbol{\lambda}}) \bullet M^{\boldsymbol{\tau\phi\lambda}}(i,k) \tag{3.20}$$

This contextual term can be viewed as a second input nonlinearity model with the model parameters $\mathbf{w}^{\boldsymbol{\tau}}$ and $\mathbf{w}^{\boldsymbol{\phi}}$ representative of relative differences in time and frequency respectively. In a similar vein to $\mathbf{w}^{\mathbf{l}}$ in equation 3.15, $\mathbf{w}^{\boldsymbol{\lambda}}$ transforms the contextual sound energy in terms of a set of P basis functions $h_p(s)$ (identical to $g_l(s)$ described earlier [1]).

Putting everything together, this fully separated context model can be expressed (in component notation) as

$$\hat{r}(i) = c + \sum_{jkl} w^{\mathbf{t}}_j w^{\mathbf{f}}_k w^{\mathbf{l}}_l M^{\mathbf{itfl}}_{ijkl} \left(c_2 + \sum_{mnp} w^{\boldsymbol{\tau}}_m w^{\boldsymbol{\phi}}_n w^{\boldsymbol{\lambda}}_p [M^{\boldsymbol{\tau\phi\lambda}}(i-j+1,k)]_{mnp} \right) \tag{3.21}$$

And again, this fully separated model can be written in multilinear form. To do so,

[1] This need not be the case in general.

we can define a final, now seven-dimensional array $\mathbf{Q}^{\mathbf{itfl}\boldsymbol{\tau}\boldsymbol{\phi}\boldsymbol{\lambda}}$ as follows

$$Q^{\mathbf{itfl}\boldsymbol{\tau}\boldsymbol{\phi}}_{ijklmnp} = \begin{cases} M^{\mathbf{itfl}}_{ijkl}[M^{\boldsymbol{\tau}\boldsymbol{\phi}\boldsymbol{\lambda}}(i-j+1,k)]_{mnp} & (j,k,l,m,n,p) \leq (J,K,L,M,N,P) \\ 1 & (j,k,l) = (J+1,K+1,L+1), \\ & (m,n,p) = (M+1,N+1,P+1) \\ M^{\mathbf{itfl}}_{ijkl} & (j,k,l) \leq (J,K,L), \\ & (m,n,p) = (M+2,N+2,P+2) \\ 0 & otherwise \end{cases} \tag{3.22}$$

Finally, with appropriate augmentation of the parameter vectors, the model can be written in its full multilinear form

$$\hat{\mathbf{r}} = (\mathbf{w^t} \otimes \mathbf{w^f} \otimes \mathbf{w^l} \otimes \mathbf{w}^{\boldsymbol{\tau}} \otimes \mathbf{w}^{\boldsymbol{\phi}} \otimes \mathbf{w}^{\boldsymbol{\lambda}}) \bullet \mathbf{Q}^{\mathbf{itfl}\boldsymbol{\tau}\boldsymbol{\phi}\boldsymbol{\lambda}} \tag{3.23}$$

At this point, it is also worth noting that since this model framework is quite generously parametrised, several choices of parameters can lead to the same global mapping. This can become somewhat problematic since the parameters within this model represent structures of particular interest that have to be interpreted. The primary degeneracy is one of scaling. That is, scaling one parameter vector by a factor n can be compensated by scaling another parameter vector by $\frac{1}{n}$. This is discussed in detail by (**?**). Typically, such a degeneracy can be handled by rescaling the constant c_2 to 1, and then rescaling each of the parameter vectors internally.

3.3.2 The Extended Context Model

Throughout the rest of this thesis, we will (almost exclusively) be working with a particular version of the context model that is slightly simpler than what has just been presented. Thus far, all contextual modulations have been modelled via the multilinear terms of $\mathbf{w}^{\boldsymbol{\tau}}$ and $\mathbf{w}^{\boldsymbol{\phi}}$, resulting in fully separable contextual interactions. Here, we will generalise this such that we have a full-rank contextual field $\mathbf{w}^{\boldsymbol{\tau}\boldsymbol{\phi}}$. This, now inseparable, field, we will refer to as the *contextual gain field* or CGF. Functionally, this now implies that frequency-difference-dependent contextual modulations can now be time-difference-dependent. Similarly, we will also utilise a full-rank principal field $\mathbf{w^{tf}}$. This, we will refer to as the *principal receptive field* or PRF. One final simplification that we will make is to disregard the original input nonlinearity flavour of the model, and restrict amplitude transformations to be linear. This will serve to further simplify analysis and to aid model estimation due to potential elimination of local optima in the objective function. The reduction in parameter count is also likely to help reduce the amount of overfitting when the model is fit (this will be discussed in more depth later).

Such an extended context model (with a fixed input nonlinearity) takes on a bilinear form, denoted by

$$\hat{\mathbf{r}} = (\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q}^{\mathbf{itf}\boldsymbol{\tau\phi}} \tag{3.24}$$

This model is shown schematically (later) in figure 4.1.

The stimulus array $\mathbf{Q}^{\mathbf{itf}\boldsymbol{\tau\phi}}$ is similar in flavour to what was defined in equation 3.22 but it now lacks the sound level components. It is fully defined as

$$Q^{\mathbf{itf}\tau\phi}_{ijkmn} = \begin{cases} M^{\mathbf{itf}}_{ijk}[M^{\tau\phi}(i,k)]_{mn} & (j,k,m,n) \leq (J,K,M,N) \\ 1 & (j,k,m,n) = (J+1,K+1,M+1,N+1) \\ M^{\mathbf{itf}}_{ijk} & (j,k) \leq (J,K), (m,n) = (M+2,N+2) \\ 0 & otherwise \end{cases} \tag{3.25}$$

Here, $\mathbf{M}^{\mathbf{itf}}$ is identical to the stimulus representation used in STRF estimation. This forms the basis for the contextual part of the stimulus representation that is defined as

$$[M^{\boldsymbol{\tau\phi}}(i,k)]_{mn} = \begin{cases} 0 & \text{if}(m,n) = (1,\Phi) \\ M^{\mathbf{itf}}_{im(k-\Phi-1+n)} & (otherwise) \end{cases} \tag{3.26}$$

3.3.3 THE SPLIT CONTEXT MODEL

In intuiting how the context model works, we have previously described the CGF component acting to modulate the sound levels within the spectrogram, prior to spectrotemporal summation through the use of the PRF. This is highly dependent upon the sign of the underlying PRF, in the sense that a positive CGF weight (for example) will always provide enhancement of whatever is present in the PRF. That is, positive (excitatory) components will be made more positive, and negative (inhibitory) components will be made more negative. Thus, for intuitive purposes, we can think of the CGF as actually modulating the values of the PRF itself. In this sense, the single CGF acts upon every individual weight within the PRF, whether it be excitatory or inhibitory, loud or soft. This leads us to define subtle variants of the context model, such that we have a multi-CGF model, whereby each CGF acts upon a different component in the PRF.

Mathematically, in order to achieve this, we want to hold $\mathbf{w^{tf}}$ constant, and fit separate $\mathbf{w}^{\boldsymbol{\tau\phi}}$s to different subsets of this underlying PRF. From a computational perspective, the PRF just consists of a set of $(J \times K)$ time-frequency pairs $\{t, f\}$, where each of these pairs denotes a different value within the field. Thus, we can write down a new rate equation for a split context model, whereby we sum over S terms, where S denotes the number of $\{t, f\}$ sets being used (which also dictates the number of CGFs). Mathematically, this can be denoted as

$$\begin{aligned}
\hat{r}(i) &= c + \\
&\sum_{\{jk\}\in\rho_1} w^{\mathbf{tf}}_{jk} M^{\mathbf{itf}}_{ijk} \left(1 + \sum_{mn} {w_{\mathbf{1}}}^{\boldsymbol{\tau\phi}}_{mn} [M^{\boldsymbol{\tau\phi}}(i-j+1,k)]_{mn}\right) + \\
&\sum_{\{jk\}\in\rho_2} w^{\mathbf{tf}}_{jk} M^{\mathbf{itf}}_{ijk} \left(1 + \sum_{mn} {w_{\mathbf{2}}}^{\boldsymbol{\tau\phi}}_{mn} [M^{\boldsymbol{\tau\phi}}(i-j+1,k)]_{mn}\right) + \\
&\vdots \\
&\sum_{\{jk\}\in\rho_s} w^{\mathbf{tf}}_{jk} M^{\mathbf{itf}}_{ijk} \left(1 + \sum_{mn} {w_{\mathbf{S}}}^{\boldsymbol{\tau\phi}}_{mn} [M^{\boldsymbol{\tau\phi}}(i-j+1,k)]_{mn}\right) \qquad (3.27)
\end{aligned}$$

where $\rho_1 \cdots \rho_s$ are sets that contain the time-frequency $\{j, k\}$ pairs of interest.

This model is shown schematically (later) in figure 4.2.

3.4 Parameter Estimation in Multilinear Models

Thus far, we have specified several different models that can be cast within a multilinear framework, but we have not discussed how the parameters within these models can be estimated. This section details how one can perform estimation within this framework.

3.4.1 Alternating Least Squares

In order to carry out estimation, we have to define a cost function (or error) and find parameters that minimise this function. A suitable choice is the squared error between the true response $\mathbf{r}$ and the response predicted under the model $\hat{\mathbf{r}}$. For the general form of a multilinear model

$$
\begin{aligned}
\mathcal{E} &= ||\mathbf{r} - \hat{\mathbf{r}}||^2 \\
&= ||\mathbf{r} - ((\mathbf{a} \otimes \mathbf{b} \otimes \cdots \otimes \mathbf{z}) \bullet \mathbf{Q})\,||^2 \qquad (3.28)
\end{aligned}
$$

Ahrens et al. (2008a) show that such a squared error can be minimised by cycling through a set of update equations, each of which resembles the solution to a classical linear regression problem (the ordinary least squares solution). This we refer to as Alternating Least Squares, or ALS. A set of such update equations corresponding to the bilinear reduced form of the context model will be derived later, in section 3.4.2. The following set of equations

$$
\begin{aligned}
\mathbf{A} &= (\mathbf{b} \otimes \mathbf{c} \cdots \otimes \mathbf{z}) \bullet \mathbf{Q} & \mathbf{a} &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{r} \\
\mathbf{B} &= (\mathbf{a} \otimes \mathbf{c} \cdots \otimes \mathbf{z}) \bullet \mathbf{Q} & \mathbf{b} &= (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{r} \\
&\vdots & &\vdots \\
\mathbf{Z} &= (\mathbf{a} \otimes \mathbf{b} \cdots \otimes \mathbf{y}) \bullet \mathbf{Q} & \mathbf{z} &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{r} \qquad (3.29)
\end{aligned}
$$

can be derived by differentiating equation 3.28 with respect to each parameter vector. These equations are applied iteratively, alternating between the different updates. Since each iteration of the algorithm will decrease the squared error and since $\mathcal{E}$ is non-negative, the iterations are guaranteed to converge to an optimum within the parameter space.

3.4.2 UPDATE EQUATIONS - STANDARD MODEL

Although the multilinear notation (through the use of high-dimensional stimulus arrays) allows us to formulate these models in a particularly elegant way, there are, of course, some computational issues. In practice, storing even a five-dimensional stimulus tensor can be particularly memory intensive. Consider, for example, the extended context model detailed in equation 3.24. If we were to consider 3000 data points ($I = 3000$), a PRF of dimension ($J = 15$) $\times$ ($K = 48$), and a CGF of dimension ($M = 13$) $\times$ ($N = 25$), which would be a standard set of units, then the full five-dimensional stimulus tensor $\mathbf{Q}^{\mathbf{itf}\boldsymbol{\tau}\boldsymbol{\phi}}$ would have to contain ($3000 \times 15 \times 48 \times 13 \times 25 = 702,000,000$) elements. This could be somewhat computationally difficult. There are, of course, ways around this problem. The stimulus that we typically utilise in our experiments is largely sparse. As a result of this, one can use a sparse representation of the stimulus spectrogram, wherein only the non-zero elements of the high-dimensional stimulus array are actually stored. This significantly reduces computational load. If one is dealing with a very dense stimulus however (such as a temporally orthogonal ripple combination (TORC; Klein et al. (2000)), or a natural sound), then a sparse representation is not appropriate since the stimulus spectrograms do not typically contain many (if any) zero entries. In these kind of cases, equivalent calculations can be carried out directly from the spectrogram itself, such that a full stimulus tensor does not have to be constructed and stored. Next, we derive the update equations to allow for estimation of $\mathbf{w}^{\mathbf{tf}}$ and $\mathbf{w}^{\boldsymbol{\tau}\boldsymbol{\phi}}$ in the extended context model of equation 3.24.

3.4.2.1 UPDATE FOR $\mathbf{w}^{\mathbf{tf}}$

In component notation, the rate equation of the reduced context model is given by

$$\hat{r}(i) = c_1 + \sum_{j=1}^{J} \sum_{k=1}^{K} w_{jk}^{\mathbf{tf}} s(i-j+1,k) \left(1 + \sum_{m=0}^{M} \sum_{n=-N}^{N} w_{mn}^{\boldsymbol{\tau}\boldsymbol{\phi}} s(i-j+1-m, k+n) \right) \tag{3.30}$$

Note that here, we have slightly altered the summation limits. This is purely for algebraic simplicity. The j and k summations are identical to before. The summations over m and n are subtly different. The m summation begins at 0 (rather than 1), in order to incorporate a contextual time-lag of 0 into the model (thus the CGF $\boldsymbol{\tau}$ dimensionality will be $(M+1)$). The subscript n corresponds to frequency deviation and thus can be either positive or negative, with N denoting the maximum allowed deviation, leading to a CGF $\boldsymbol{\phi}$ dimensionality of $(2 \times N + 1)$.

We then multiply out the brackets to construct a new three-dimensional array A_{ijk} such that

$$A_{ijk} = s(i-j+1,k) + s(i-j+1,k) \sum_{mn} w_{mn}^{\boldsymbol{\tau}\boldsymbol{\phi}} s(i-j+1-m, k+n) \tag{3.31}$$

This reduces equation 3.30 to

$$\hat{r}(i) = c_1 + \sum_{jk} w_{jk}^{\mathbf{tf}} A_{ijk} \tag{3.32}$$

Thus, in order to update $\mathbf{w^{tf}}$, we hold $\mathbf{w^{\tau\phi}}$ fixed and regress using

$$\hat{\mathbf{r}} = \mathbf{c_1} + \mathbf{w^{tf}} \bullet \mathbf{A^{itf}} \tag{3.33}$$

3.4.2.2 Update for $\mathbf{w^{\tau\phi}}$

In order to derive the update equation for $\mathbf{w^{\tau\phi}}$ we start by rewriting equation 3.30 as

$$\hat{r}(i) = c_1 + \sum_{jk} w^{\mathbf{tf}}_{jk} s(i-j+1,k) + \sum_{mn} w^{\boldsymbol{\tau\phi}}_{mn} \sum_{jk} w^{\mathbf{tf}}_{jk} s(i-j+1,k) s(i-j+1-m,k+n) \tag{3.34}$$

We then simply bring one term across, such that we can then regress against an augmented firing rate vector

$$\begin{aligned} \hat{r}(i) - \sum_{jk} w^{\mathbf{tf}}_{jk} s(i-j+1,k) &= c_1 + \sum_{mn} w^{\boldsymbol{\tau\phi}}_{mn} \times \\ & \quad \sum_{jk} \mathbf{w}^{\mathbf{tf}}_{jk} s(i-j+1,k) s(i-j+1-m,k+n) \\ \hat{r}(i) - \sum_{jk} w^{\mathbf{tf}}_{jk} s(i-j+1,k) &= c_1 + \sum_{mn} w^{\boldsymbol{\tau\phi}}_{mn} B_{imn} \end{aligned} \tag{3.35}$$

Thus, in order to update $\mathbf{w^{\tau\phi}}$, we hold $\mathbf{w^{tf}}$ fixed and regress using

$$\hat{\mathbf{r}} - \mathbf{w^{tf}} \bullet \mathbf{M^{itf}} = \mathbf{w}^{\tau\phi} \bullet \mathbf{B^{i\tau\phi}} \tag{3.36}$$

where $M^{\mathbf{itf}}_{ijk} = s(i-j+1,k)$.

3.4.3 Update Equations - Split Model

For completeness, we also provide the update equations for the split model that was discussed in section 3.3.3. Here, for simplicity, we focus on a model with two CGFs.

3.4.3.1 Update for $\mathbf{w^{tf}}$

We can expand the brackets in the same way as in section 3.4.2.1, in order to construct the stimulus tensors $\mathbf{A_1}$ and $\mathbf{A_2}$. That is

$$\hat{r}(i) = c_1 + \sum_{jk} \mathbf{1_1} w^{\mathbf{tf}}_{jk} A_{\mathbf{1}\,ijk} + \sum_{jk} \mathbf{1_2} w^{\mathbf{tf}}_{jk} A_{\mathbf{2}\,ijk} \tag{3.37}$$

Here, we utilise indicator variables that will act to set $\{t, f\}$ elements in $\mathbf{w^{tf}}$, $\mathbf{A_1^{itf}}$ and $\mathbf{A_2^{itf}}$ to 0 if they are not present within the correct set:

$$\mathbf{1_s} = \begin{cases} 1 & \text{if} \quad \{j,k\} \in \rho_s \\ 0 & otherwise \end{cases} \tag{3.38}$$

Thus, we can write

$$\hat{\mathbf{r}} = c_1 + \mathbf{w^{tf}} \bullet \left(\mathbf{A_1^{itf}} + \mathbf{A_2^{itf}}\right) \tag{3.39}$$

which we can regress in the usual way to estimate $\mathbf{w^{tf}}$.

3.4.3.2 Update for $\mathbf{w_1^{\tau\phi}}$and $\mathbf{w_2^{\tau\phi}}$

We can also update $\mathbf{w_1^{\tau\phi}}$ and $\mathbf{w_2^{\tau\phi}}$ in a similar way. Following equation 3.35, we end up with

$$\hat{r}(i) - \sum_{jk} w_{jk}^{\mathbf{tf}} s(i-j+1,k) = c_1 + \sum_{mn} w_{\mathbf{1}}{}_{mn}^{\boldsymbol{\tau\phi}} B_{\mathbf{1}\,imn} + \sum_{mn} w_{\mathbf{2}}{}_{mn}^{\boldsymbol{\tau\phi}} B_{\mathbf{2}\,imn} \tag{3.40}$$

Holding $\mathbf{w^{tf}}$ fixed, we can then simply regress in the usual way to update $\mathbf{w_1^{\tau\phi}}$ and $\mathbf{w_2^{\tau\phi}}$.

3.4.4 Control of Overfitting

A particular concern when dealing with such a large number of parameters is the problem of overfitting. Some of the models that we wish to estimate contain upwards of 1000 parameters, which can result in the incorrect "explanation" of noise within the data. This kind of behaviour can be discouraged by utilising some form of statistical regularisation.

Here, we will adopt a Bayesian perspective which provides a particularly useful way of performing such regularisation by supplying prior information (in the form of regularisation parameters) about the model parameters themselves. The least squares solution for a regularised linear regression problem takes the form

$$\mathbf{a} = (\mathbf{A}^T\mathbf{A} + \sigma^2\mathbf{C})^{-1}\mathbf{A}^T\mathbf{r} \tag{3.41}$$

where the covariance matrix C contains information regarding our prior beliefs about the parameters.

In order to obtain appropriate regularisation matrices, we follow the work of Sahani and Linden (2003a) who develop a method to adapt the covariance structure of the parameters to the *evidence* given by the data. The method that they propose effectively controls the spectral and temporal smoothness between parameters and was therefore termed Automatic Smoothness Determination (ASD).

We first develop how evidence is defined in this context, and then specify details of the covariance matrix. The following derivation follows that of Sahani and Linden (2003a).

From a probabilistic perspective, the squared-error term $\mathcal{E}$ (equation 3.28) corresponds to a Gaussian likelihood. Specifically, the least-squares solution to any regression problem is identical to the maximum likelihood (ML) value of the parameter vector $\mathbf{w}$ for a probabilistic regression model with Gaussian noise of constant variance σ^2

$$r_t|\mathbf{x}_t \quad \sim \quad \mathcal{N}(\mathbf{w}^T\mathbf{x}_t, \sigma^2) \tag{3.42}$$

For consistency with Sahani and Linden (2003a), we now describe the input as a matrix X, the t^{th} column of which is the input lag-vector $\mathbf{x}_t$ (a lagged representation of the stimulus spectrogram). The outputs are denoted as a row vector $\mathbf{r}$, the t^{th} element of which is r_t.

We can write down the Gaussian likelihood as

$$P(\mathbf{r}|X, \mathbf{w}, \sigma^2) \approx \exp\left(-\frac{1}{2}\frac{(\mathbf{r}-\mathbf{w}^TX)(\mathbf{r}-\mathbf{w}^TX)^T}{\sigma^2}\right) \tag{3.43}$$

We can then obtain the joint density of $\mathbf{r}$ and $\mathbf{w}$ by multiplication with a Gaussian prior of zero mean (since we have no prior reason to favour either positive or negative weights) and a covariance matrix C

$$P(\mathbf{r}, \mathbf{w}|X, C, \sigma^2) \approx \exp\left(-\frac{1}{2}\left(\frac{(\mathbf{r}-\mathbf{w}^TX)(\mathbf{r}-\mathbf{w}^TX)^T}{\sigma^2} - \mathbf{w}^TC^{-1}\mathbf{w}\right)\right) \tag{3.44}$$

Since the likelihood and the prior are conjugate, the posterior distribution on $\mathbf{w}$ is also Gaussian with variance $\Sigma = \left(\frac{XX^T}{\sigma^2} + C^{-1}\right)^{-1}$ and mean $\mu = \Sigma\frac{X\mathbf{r}^T}{\sigma^2}$. Integrating out the parameters $\mathbf{w}$ then gives us an expression for the evidence (or marginal likelihood) of the data

$$P(\mathbf{r}|X, C, \sigma^2) \approx \exp\left(-\frac{\mathbf{r}}{2}\left(\frac{I}{\sigma^2} - \frac{X^T \Sigma X}{\sigma^4}\right)\mathbf{r}^T\right) \tag{3.45}$$

Finally, differentiating this expression with respect to a parameter θ that parametrises the covariance matrix C we get

$$\frac{\partial}{\partial \theta} \log P(\mathbf{r}|X, C, \sigma^2) = \frac{1}{2} \operatorname{Tr}\left((C - \Sigma - \mu\mu^T)\frac{\partial}{\partial \theta} C^{-1}\right) \tag{3.46}$$

This powerful framework for performing evidence optimisation allows one to carry out hyperparameter optimisation on the parameters within any chosen covariance matrix. One such covariance matrix underlies the ASD algorithm of Sahani and Linden (2003a). They define

$$C = \exp\left(-\rho - \frac{1}{2}\left(\frac{\triangle_s}{\delta_s^2} + \frac{\triangle_t}{\delta_t^2}\right)\right) \tag{3.47}$$

where $\triangle_s$ and $\triangle_t$ are distance matrices, wherein the $(i, j)^{th}$ element of each gives the squared distance between the weights w_i and w_j in terms frequency and time respectively. As a result of these squared distances, the free parameters within the covariance matrix (δ_s and δ_t) set the correlation distances for the weights along the spectral and temporal dimensions. Thus, large values of either of these hyperparameters will favour smoothness in the relevant dimension. Estimation of these parameters simply amounts to gradient descent through the use of equation 3.46.

This ASD algorithm can be easily (but sub-optimally) integrated into the ALS framework for parameter estimation in multilinear models. Care must be taken however, to not use ASD updates at every iteration of the algorithm. This is simply because the covariance hyperparameters (that govern the smoothness) will change between iterations, which means that convergence of the fitting procedure is no longer guaranteed. As a result, ASD updates are best utilised for a small number of iterations, in order to achieve a reasonable estimate of the spectrotemporal smoothness of a given unit. Once the hyperparameters are fixed, the remaining iterations are again guaranteed to converge.

3.4.5 Variational Approximations to Bilinear Systems

Alternating least squares (although guaranteed to converge) can, on occasion, lead to erroneous receptive field structure, especially when used in combination with ASD. Unfortunately, this is not a principled way of handling estimation in the multilinear setting due to the fact that the smoothness parameters are estimated separately per parameter vector, as opposed to jointly for the entire system. From a Bayesian standpoint this is somewhat less than ideal, since uncertainty about one parameter vector is not taken into account in the estimation of another.

Here we will present a principled empirical Bayesian approach such that propagation of uncertainty can be correctly handled.

As before, the full-rank bilinear model is given by

$$\hat{\mathbf{r}} = (\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q}^{\mathbf{itf}\boldsymbol{\tau\phi}} \tag{3.48}$$

For convenience, we will make another small change to notation. As it stands, the stimulus array ($\mathbf{Q}^{\mathbf{itf}\boldsymbol{\tau\phi}}$) is five-dimensional. After concatenation of indices in the different fields however, the array becomes three-dimensional; $(j,k) \to a = 1..a_{\text{max}}$, $(m,n) \to b = 1..b_{max}$. Thus we use $w_a^{\mathbf{tf}}$, $w_b^{\boldsymbol{\tau\phi}}$, and Q_{iab}.

3.4.5.1 Automatic Regularisation

Rather than specifying an algorithm to perform ASD for the bilinear system, we present a slightly different type of regularisation for reasons of stability (fewer ill-conditioned matrix inversions). The crucial difference with this approach is that rather than specifying prior *covariance* matrices, we will specify the regularisation through the use of the *inverse covariance* (or *precision*) matrix. Precision matrices can be directly interpreted as a cost on the parameter vectors, because of the way in which they appear within the squared-error cost function. For a linear system we have

$$\mathcal{E} = \frac{1}{2\sigma^2}||\mathbf{r} - \hat{\mathbf{r}}||^2 + \mathbf{w}^T D\mathbf{w} \tag{3.49}$$

How does one choose a suitable D? A common choice is to penalise derivatives. For example, if one wanted to penalise the first derivative, a D could be chosen such that $w^T D w = \alpha \sum_j (w_{j+1} - w_j)^2$ (with α being the parameter controlling the degree of regularisation). This is satisfied by choosing D to be a matrix with 2's on the diagonal and -1's just off the diagonal. In general, choosing Z to be a differentiating matrix

$$Z = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.50}$$

to penalise the first derivative, $D_1 = Z^T Z$. In order to penalise the second derivative, one would choose $D_2 = D_1^T D_1$. Moreover, to penalize large values of $\mathbf{w}$, $D_0 = I$ could be chosen (which is equivalent to ridge regression). For our purposes, however, we choose to use a linear combination of ridge, first and second derivatives, such that our precision matrix is denoted thus

$$D = \alpha_0 D_0 + \alpha_1 D_1 + \alpha_2 D_2 \tag{3.51}$$

Finally, in order to make this applicable for two-dimensional receptive fields, we must use separate smoothing in both dimensions, which yields five parts to the sum above.

3.4.5.2 Evidence Optimisation via Variational Approximation

Disclaimer. What follows is an algorithm for a rigorous treatment of evidence approximation in a bilinear model. It is, however, incredibly computationally intensive. As a result, it is presented here for mathematical completeness. The estimation algorithm we utilise in the coming chapters, is essentially a simplified version of the ALS algorithm. From a qualitative perspective, the differences between the two approaches are relatively minimal (in regards to the structure of the estimated receptive fields).

In section 3.4.4, we discussed how the ASD algorithm was based upon optimising the evidence of the model. In the simple linear case that we presented, everything was conjugate and thus the evidence (the probability of the data given the regularisation parameters, with the model parameters integrated out) was tractable. The ultimate goal is to be able to maximise this evidence with respect to a given regularisation parameter (such as a smoothness parameter) in order to establish the optimal smoothness of the corresponding model parameter. Due to the tractability of the linear example, the evidence could be written down explicitly in closed form, and the maximisation could be carried out numerically by taking the relevant derivatives and performing gradient ascent. The full bilinear (or multilinear) evidence optimisation is intractable. Thus, here we develop a method such the evidence can be adequately approximated, and correct uncertainty propagation can be handled.

We choose to use a variational approximation (Jordan et al., 1999; Beal, 2003), utilising a factorised distribution $q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}})q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}})$ over the parameters to obtain a lower bound on the evidence to which a standard expectation maximisation (EM; Dempster et al. (1977)) optimisation can be applied. For brevity, only the gist of the algorithm is presented here. The full derivation can be found in the appendix, in section 3.5.

The (log) evidence of the bilinear model is given by

$$\log P(\mathbf{r}|\hat{\sigma}^2, \boldsymbol{\alpha}^{\mathbf{tf}}, \boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) = \log \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \tag{3.52}$$

We then make use of the factorised distribution above, in order to lower bound this evidence

$$\begin{aligned}
&\log P(\mathbf{r}|\hat{\sigma}^2, \boldsymbol{\alpha}^{\mathbf{tf}}, \boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \\
&\quad = \log \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} \frac{q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}})q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}})}{q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}})q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}})} P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \\
&\quad \geq \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log\left(P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}})\right) \\
&\qquad\quad - \int d\mathbf{w}^{\mathbf{tf}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) \log\left(q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}})\right) - \int d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log\left(q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}})\right) \\
&\quad = \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log\left(P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}})\right) \\
&\qquad + H(q^{\mathbf{tf}}) + H(q^{\boldsymbol{\tau\phi}}) \\
&\quad \equiv \mathcal{F}(q^{\mathbf{tf}}, q^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2, \boldsymbol{\alpha}^{\mathbf{tf}}, \boldsymbol{\alpha}^{\boldsymbol{\tau\phi}})
\end{aligned} \tag{3.53}$$

Here $\mathcal{F}$ is the free energy, and the two H terms are entropies. We use the EM algorithm to optimise the former quantity.

3.4.5.3 E Step

Goal: optimise $\mathcal{F}$ w.r.t. $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$ (by taking variational derivatives w.r.t. $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$ and setting these to zero).

Differentiating and setting equal to zero yields

$$q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) = \exp(\lambda - 1) \exp\left\langle \log\left(P(\hat{\mathbf{r}}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2)\right)\right\rangle_{q^{\boldsymbol{\tau\phi}}} P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) \tag{3.54}$$

with the update for $q^{\boldsymbol{\tau\phi}}$ being found in a similar fashion.

It should be reasonably easy to see that $\log P(\hat{\mathbf{r}}|\mathbf{w^{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2)$ is quadratic in $\mathbf{w^{tf}}$ (and this gets preserved in the average $\langle\cdots\rangle_{q^{\boldsymbol{\tau\phi}}}$). Thus, if the prior $P(\mathbf{w^{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}})$ is Gaussian, then $q^{\mathbf{tf}}$ is also Gaussian. We can give the parameters of these Gaussians a name: let $q^{\mathbf{tf}} = \mathcal{N}(\boldsymbol{\mu}^{\mathbf{tf}}, \boldsymbol{\Sigma}^{\mathbf{tf}})$, and $q^{\boldsymbol{\tau\phi}} = \mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}, \boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}})$. The important thing to note here is that $\boldsymbol{\mu}^{\mathbf{tf}}$ and $\boldsymbol{\Sigma}^{\mathbf{tf}}$ depend on $\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}}$ (and vice-versa). As such, updating these parameters allows for the correct propagation of uncertainty. Deriving these update equations is not trivial, and thus the details can be found within the appendix, in section 3.5.

3.4.5.4 M Step

Goal: with $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$ held fixed, $\mathcal{F}$ is maximised with respect to the hyperparameters $\boldsymbol{\alpha}^{\mathbf{tf}}$,$\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}$, and the noise scale $\hat{\sigma}^2$. The noise scale $\hat{\sigma}^2$ has an exact update, but the hyperparameter optimisation needs to be done through the use of gradient ascent, since there is no closed-form solution. The gradient is obtained by differentiating $\mathcal{F}$ w.r.t. $\boldsymbol{\alpha}^{\mathbf{tf}}$ and $\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}$.

3.4.6 Full Algorithm

With the key elements in place, the full algorithm can then be defined as:

1. The algorithm is iterative, so first assume that we have values for $\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}}$.

2. Update the values of $\boldsymbol{\mu}^{\mathbf{tf}}$ and $\boldsymbol{\Sigma}^{\mathbf{tf}}$.

3. Update the values of $\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}}$.

4. Update the noise scale $\hat{\sigma}^2$ exactly, as a closed-form solution exists.

5. Perform gradient ascent on the function $\mathcal{F}$, with respect to the parameters $\boldsymbol{\alpha}^{\mathbf{tf}}$ and $\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}$, as no closed-form solution exists.

6. Repeat steps 2-5 until all parameters have converged.

3.5 Appendix A: Variational EM for Bilinear Systems

In section 3.4.5.2 we detailed an empirical Bayesian algorithm for performing approximate evidence optimisation in a bilinear system. For brevity, only the key results were presented. The full derivation of the variational EM algorithm follows in this appendix.

The (log) evidence of the bilinear model is

$$\log P(\mathbf{r}|\hat{\sigma}^2, \boldsymbol{\alpha}^{\mathbf{tf}}, \boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) = \log \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\alpha^{\boldsymbol{\tau\phi}}) \tag{3.55}$$

We then make use of a factorised distribution $q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}})$, in order to lower bound this evidence

$$\begin{aligned}
&\log P(\mathbf{r}|\hat{\sigma}^2, \boldsymbol{\alpha}^{\mathbf{tf}}, \boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \\
&\quad = \log \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \\
&\quad = \log \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} \frac{q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}})}{q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}})} P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \\
&\quad \geq \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log \left(\frac{P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}})}{q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}} (\mathbf{w}^{\boldsymbol{\tau\phi}}) \right) \\
&\quad = \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log \left(P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \right) \\
&\qquad - \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log \left(q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \right) \\
&\quad = \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log \left(P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \right) \\
&\qquad - \int d\mathbf{w}^{\mathbf{tf}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) \log \left(q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) \right) - \int d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log \left(q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \right) \\
&\quad = \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log \left(P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}) \right) \\
&\qquad + H(q^{\mathbf{tf}}) + H(q^{\boldsymbol{\tau\phi}}) \\
&\quad \equiv \mathcal{F}(q^{\mathbf{tf}}, q^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2, \boldsymbol{\alpha}^{\mathbf{tf}}, \boldsymbol{\alpha}^{\boldsymbol{\tau\phi}})
\end{aligned} \tag{3.56}$$

Here $\mathcal{F}$ is the free energy, and the two H terms are entropies. We use the EM algorithm to optimise the former quantity.

3.5.1 E Step

Goal: optimise $\mathcal{F}$ w.r.t. $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$ (by taking variational derivatives w.r.t. $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$ and setting these to zero).

The derivation that follows will focus on $q^{\mathbf{tf}}$ only. $q^{\boldsymbol{\tau\phi}}$ follows in a very similar fashion. We start by adding a Lagrange multiplier to $\mathcal{F}$ to constrain $q^{\mathbf{tf}}$ to be normalised

$$\begin{aligned}
&\frac{\delta\left(\mathcal{F} + \lambda(\int d\mathbf{w}^{\mathbf{tf}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) - 1)\right)}{\delta q^{\mathbf{tf}}} \\
&\quad = \int d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log\left(P(\hat{\mathbf{r}}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2) P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}})\right) \\
&\qquad - \log(q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) - 1 + \lambda
\end{aligned} \tag{3.57}$$

Setting this to zero yields

$$q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) = \exp(\lambda - 1) \exp\left\langle \log\left(P(\hat{\mathbf{r}}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2)\right)\right\rangle_{q^{\boldsymbol{\tau\phi}}} P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) \tag{3.58}$$

(note the $P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}})$ is also inside an $\exp\langle\log(...)\rangle_{q^{\boldsymbol{\tau\phi}}}$ but since it's independent of $\mathbf{w}^{\boldsymbol{\tau\phi}}$ this just becomes $P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}})$.) Here λ is a constant that serves to normalize $q^{\mathbf{tf}}$.

Now $\log P(\hat{\mathbf{r}}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2)$ is quadratic in $\mathbf{w}^{\mathbf{tf}}$ and this gets preserved in the average $\langle \ldots \rangle_{q^{\boldsymbol{\tau\phi}}}$, so that if the prior $P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}})$ is Gaussian, then $q^{\mathbf{tf}}$ is also Gaussian. We can now give the parameters of these Gaussians a name: let $q^{\mathbf{tf}} = \mathcal{N}(\boldsymbol{\mu}^{\mathbf{tf}}, \boldsymbol{\Sigma}^{\mathbf{tf}})$ and $q^{\boldsymbol{\tau\phi}} = \mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}, \boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}})$. The algorithm is iterative, so we can assume we have values for $\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}}$ and use these to update the values of $\boldsymbol{\mu}^{\mathbf{tf}}$ and $\boldsymbol{\Sigma}^{\mathbf{tf}}$. To find these, we first ignore the prior and look at the likelihood term (absorbing the log-determinant in the constant),

$$\begin{aligned}
&\left\langle \log\left(P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}, \hat{\sigma}^2)\right)\right\rangle_{q^{\boldsymbol{\tau\phi}}} \\
&\quad = \text{const} - \frac{1}{2\hat{\sigma}^2}\left\langle (\mathbf{r} - (\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q})^2 \right\rangle_{q^{\boldsymbol{\tau\phi}}} \\
&\quad = \text{const} - \frac{1}{2\hat{\sigma}^2}\langle \mathbf{r}^2 - 2\mathbf{r}^T\left((\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q}\right) \\
&\qquad + \left((\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q}\right)^T \left((\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q}\right)\rangle_{q^{\boldsymbol{\tau\phi}}}
\end{aligned} \tag{3.59}$$

We can now simplify the quadratic term. Ignoring the $\frac{1}{2\hat{\sigma}^2}$ for the moment and writing in component notation, the quadratic term is

$$\left\langle \sum_{iaa'bb'} w_a^{\mathbf{tf}} w_{a'}^{\mathbf{tf}} w_b^{\boldsymbol{\tau\phi}} w_{b'}^{\boldsymbol{\tau\phi}} Q_{iab} Q_{ia'b'} \right\rangle_{q^{\boldsymbol{\tau\phi}}} = \sum_{iaa'bb'} w_a^{\mathbf{tf}} w_{a'}^{\mathbf{tf}} \left\langle w_b^{\boldsymbol{\tau\phi}} w_{b'}^{\boldsymbol{\tau\phi}} \right\rangle_{q^{\boldsymbol{\tau\phi}}} Q_{iab} Q_{ia'b'} \tag{3.60}$$

The term in angle brackets is the $(b,b')^{\text{th}}$ element of $\left\langle \mathbf{w}^{\boldsymbol{\tau\phi}} \mathbf{w}^{\boldsymbol{\tau\phi}T} \right\rangle_{q^{\boldsymbol{\tau\phi}}}$, which equals $\boldsymbol{\mu}^{\boldsymbol{\tau\phi}} \boldsymbol{\mu}^{\boldsymbol{\tau\phi}T} + \boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}}$. We shrink this expression, and reintroduce $\frac{1}{\hat{\sigma}^2}$, by setting

$$C^{\mathbf{tf}}_{aa'} = \frac{1}{\hat{\sigma}^2} \sum_{ibb'} \left(\mu^{\boldsymbol{\tau\phi}}_b \mu^{\boldsymbol{\tau\phi}}_{b'} + \Sigma^{\boldsymbol{\tau\phi}}_{bb'} \right) Q_{iab} Q_{ia'b'} \tag{3.61}$$

to get the quadratic term

$$\sum_{aa'} w^{\mathbf{tf}}_a w^{\mathbf{tf}}_{a'} C^{\mathbf{tf}}_{aa'} = \mathbf{w}^{\mathbf{tf}T} \mathbf{C}^{\mathbf{tf}} \mathbf{w}^{\mathbf{tf}} \tag{3.62}$$

In order to establish the covariance of $q^{\mathbf{tf}}$, we simply need to multiply by the prior, or equivalently, add the log-prior to the above expression. The log-prior is

$$\log P(\mathbf{w}^{\mathbf{tf}} | \boldsymbol{\alpha}^{\mathbf{tf}}) = -\frac{1}{2} \mathbf{w}^{\mathbf{tf}T} D^{\mathbf{tf}} \mathbf{w}^{\mathbf{tf}} \tag{3.63}$$

therefore the entire quadratic term is $-\frac{1}{2} \mathbf{w}^{\mathbf{tf}T} \left(C^{\mathbf{tf}} + D^{\mathbf{tf}} \right) \mathbf{w}^{\mathbf{tf}}$ and so the covariance of $q^{\mathbf{tf}}$ is

$$\boldsymbol{\Sigma}^{\mathbf{tf}} = \left(C^{\mathbf{tf}} + D^{\mathbf{tf}} \right)^{-1} \tag{3.64}$$

The mean of $q^{\mathbf{tf}}$ is found similarly and is defined via the vector

$$v^{\mathbf{tf}}_a = \sum_i r_i Q_{iab} \mu^{\boldsymbol{\tau\phi}}_b \tag{3.65}$$

and equals

$$\boldsymbol{\mu}^{\mathbf{tf}} = \frac{1}{\hat{\sigma}^2} \boldsymbol{\Sigma}^{\mathbf{tf}} v^{\mathbf{tf}} \tag{3.66}$$

This is the E-step update for $q^{\mathbf{tf}}$. The update for $q^{\boldsymbol{\tau\phi}}$ is exactly the same except the averaging is over $q^{\mathbf{tf}}$, and $\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}}$ get defined in terms of $\boldsymbol{\mu}^{\mathbf{tf}}$ and $\boldsymbol{\Sigma}^{\mathbf{tf}}$.

3.5.2 M Step

Goal: with $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$ held fixed, $\mathcal{F}$ is maximised with respect to the hyperparameters $\boldsymbol{\alpha}^{\mathbf{tf}}$,$\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}$, and the noise scale $\hat{\sigma}^2$. The noise scale $\hat{\sigma}^2$ has an exact update, but the hyperparameter optimisation needs to be done through the use of gradient ascent, since there is no closed-form solution. The gradient is obtained by differentiating $\mathcal{F}$ w.r.t. $\boldsymbol{\alpha}^{\mathbf{tf}}$ and $\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}$.

Since we are differentiating with respect to the hyperparameters, only the terms in $\mathcal{F}$ that depend on $\boldsymbol{\alpha}^{\mathbf{tf}}$ and $\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}$ need to be considered; namely, the log-priors averaged over the $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$. Note that the hyperparameters also appear implicitly in the distributions $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$, but since these are considered fixed in the M step of the EM algorithm, they do not contribute here. For notational purposes, we call the α-dependent terms $\bar{\mathcal{F}}$ and thus

$$\begin{aligned}
\bar{\mathcal{F}} &= \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log \left(P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}}) P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}})\right) \\
&= \left\langle \log P(\mathbf{w}^{\mathbf{tf}}|\boldsymbol{\alpha}^{\mathbf{tf}})\right\rangle_{q^{\mathbf{tf}}} + \left\langle \log P(\mathbf{w}^{\boldsymbol{\tau\phi}}|\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}})\right\rangle_{q^{\boldsymbol{\tau\phi}}} \\
&= \frac{1}{2}\log\det D^{\mathbf{tf}} + \frac{1}{2}\log\det D^{\boldsymbol{\tau\phi}} - \frac{1}{2}\text{trace}\left((\boldsymbol{\mu}^{\mathbf{tf}}\boldsymbol{\mu}^{\mathbf{tf}^T} + \boldsymbol{\Sigma}^{\mathbf{tf}})D^{\mathbf{tf}}\right) \\
&\quad -\frac{1}{2}\text{trace}\left((\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}\boldsymbol{\mu}^{\boldsymbol{\tau\phi}^T} + \boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}})D^{\boldsymbol{\tau\phi}}\right) \qquad (3.67)
\end{aligned}$$

(using $\det X^{-1} = \frac{1}{\det X}$ for the first two terms.) Now we differentiate w.r.t. the parameters that specify the D's, e.g. $\alpha_i^{\mathbf{tf}}$. The derivative can be taken inside the trace, and for the log terms, we use $\frac{\partial}{\partial\theta}\log\det X(\theta) = \mathrm{Tr}\left(X^{-1}\frac{\partial}{\partial\theta}X\right)$. Thus

$$\frac{\partial}{\partial\alpha_i^{\mathbf{tf}}}\bar{\mathcal{F}} = \frac{1}{2}\,\mathrm{Tr}\left(D^{\mathbf{tf}^{-1}}\frac{\partial}{\partial\alpha_i^{\mathbf{tf}}}D^{\mathbf{tf}}\right) - \frac{1}{2}\,\mathrm{Tr}\left((\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}\boldsymbol{\mu}^{\boldsymbol{\tau\phi}^T} + \boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}})\frac{\partial}{\partial\alpha_i^{\mathbf{tf}}}D^{\mathbf{tf}}\right) \qquad (3.68)$$

With the form of $D^{\mathbf{tf}}$ above, i.e. $D^{\mathbf{tf}}(\boldsymbol{\alpha}^{\mathbf{tf}}) = \sum_i \alpha_i^{\mathbf{tf}} D_i^{\mathbf{tf}}$, this becomes

$$\frac{\partial}{\partial\alpha_i^{\mathbf{tf}}}\bar{\mathcal{F}} = \frac{1}{2}\,\mathrm{Tr}\left(D_i^{\mathbf{tf}}\left(D^{\mathbf{tf}^{-1}} - (\boldsymbol{\mu}^{\boldsymbol{\tau\phi}}\boldsymbol{\mu}^{\boldsymbol{\tau\phi}^T} + \boldsymbol{\Sigma}^{\boldsymbol{\tau\phi}})\right)\right) \qquad (3.69)$$

The gradients over all ${\boldsymbol{\alpha}^{\mathbf{tf}}}_i$'s and all ${\boldsymbol{\alpha}^{\boldsymbol{\tau\phi}}}_i$'s can then be used in a gradient ascent to maximize $\bar{\mathcal{F}}$ (equivalently to maximize $\mathcal{F}$).

The noise scale $\hat{\sigma}^2$ is also considered a hyperparameter, and must be optimized during the M step. This parameter is found in the likelihood term, and implicitly in the Σ's and μ's. Again we consider the Σ's and μ's fixed because $q^{\mathbf{tf}}$ and $q^{\boldsymbol{\tau\phi}}$ are fixed during the M step of EM. Thus, only the likelihood term contributes here, and we define $\hat{\mathcal{F}}$ to be the part of $\mathcal{F}$ containing $\hat{\sigma}^2$. Thus

$$\begin{aligned}
\hat{\mathcal{F}} &= \int d\mathbf{w}^{\mathbf{tf}} d\mathbf{w}^{\boldsymbol{\tau\phi}} q^{\mathbf{tf}}(\mathbf{w}^{\mathbf{tf}}) q^{\boldsymbol{\tau\phi}}(\mathbf{w}^{\boldsymbol{\tau\phi}}) \log P(\mathbf{r}|\mathbf{w}^{\mathbf{tf}}, \mathbf{w}^{\boldsymbol{\tau\phi}}\}, \hat{\sigma}^2) \\
&= \text{const} - \frac{T}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\left\langle\left(\mathbf{r} - (\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q}\right)^2\right\rangle_{q^{\mathbf{tf}}, q^{\boldsymbol{\tau\phi}}} \qquad (3.70)
\end{aligned}$$

Differentiating w.r.t. $\hat{\sigma}^2$ yields

$$\frac{\partial}{\partial \hat{\sigma}^2}\hat{\mathcal{F}} = -\frac{T}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\left\langle \left(\mathbf{r} - (\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q}\right)^2 \right\rangle_{q^{\mathbf{tf}}, q^{\boldsymbol{\tau\phi}}} \tag{3.71}$$

The optimum lies at $\frac{\partial}{\partial \hat{\sigma}^2}\hat{\mathcal{F}} = 0$, which can be solved as

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{T}\left\langle \left(\mathbf{r} - (\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\boldsymbol{\tau\phi}}) \bullet \mathbf{Q}\right)^2 \right\rangle_{q^{\mathbf{tf}}, q^{\boldsymbol{\tau\phi}}} \\ &= \frac{1}{T}\left(\mathbf{r}^T\mathbf{r} - 2t_1 + t_2\right) \end{aligned} \tag{3.72}$$

where T is the number of time points and

$$\begin{aligned} t_1 &= \mathbf{r}^T\left(\boldsymbol{\mu}^{\mathbf{tf}} \otimes \boldsymbol{\mu}^{\boldsymbol{\tau\phi}}\right) \bullet \mathbf{Q} = \sum_{iab} r_i \mu_a^{\mathbf{tf}} \mu_b^{\boldsymbol{\tau\phi}} Q_{iab} \\ t_2 &= \sum_{iaa'bb'} \left(\mu_a^{\mathbf{tf}}\mu_{a'}^{\mathbf{tf}} + \Sigma_{aa'}\right)\left(\mu_b^{\boldsymbol{\tau\phi}}\mu_{b'}^{\boldsymbol{\tau\phi}} + \Sigma_{bb'}^{\boldsymbol{\tau\phi}}\right) Q_{iab} Q_{ia'b'} \end{aligned} \tag{3.73}$$

IV

Near-Simultaneous and Delayed Contextual Effects in the Mouse Thalmocortical Pathway

Outline

This chapter is the first of two primary results chapters within this thesis. It provides an application of the multilinear context model framework to neural responses in the auditory cortex and thalamus. We show that such a framework is capable of the successful estimation of nonlinear interactions from the neural responses to complex sounds, thus extending our existing knowledge of sound processing within the thalamocortical pathway. Although this chapter relies heavily on the modelling framework discussed at length in the previous chapter, it is written as a standalone piece of work, and can be read without prior knowledge of the detailed mathematics.

4.1 Introduction

Neuronal responses in the auditory cortex can be strongly and non-linearly modulated by stimulus context (Brosch et al., 1999; Brosch and Schreiner, 2000; Bartlett and Wang, 2005; Calford and Semple, 1995; Sadagopan and Wang, 2009; Bar-Yosef et al., 2002; Bar-Yosef and Nelken, 2007). As a result of this, standard linear descriptions of neuronal stimulus-response functions (i.e., spectrotemporal receptive fields (STRFs)), are not sufficient to explain auditory cortical responses to spectrally rich, temporally complex sounds. The effect of (short-term) stimulus context in the auditory thalamus (one synapse upstream of the auditory cortex) is currently not well understood (but see Wehr and Zador (2005) for an intracellular example). Ahrens et al. (2008a) introduced multilinear "context" models, which capture neuron-specific nonlinear effects of stimulus context on spiking responses to complex sounds. In such a framework, contextual effects are interpreted as non-linear stimulus interactions that modulate the input to a subsequent STRF-like linear filter. It was previously demonstrated that such context models predict rodent auditory cortical responses to complex sounds more accurately than do standard STRF models, leading to the conclusion that nonlinear contextual interactions play an important role in the cortical processing of complex sounds. The form of these contextual interactions will constitute the primary focus of this chapter.

The analysis of Ahrens et al. (2008a) assumed that the effects of stimulus context were fully separable (i.e., independent in frequency and time). Here, we use an extended context model to test this assumption, and we demonstrate that the non-linear effects of stimulus context are, in fact, largely inseparable, and fundamentally different for near-simultaneous and delayed non-simultaneous sound energy. In two populations of neurons, recorded from the mouse auditory cortex and from the auditory thalamus, we show that simultaneous sound energy provides a nonlinear positive (amplifying) gain to the subsequent linear filter, while non-simultaneous sound energy provides a negative (dampening) gain.

We also demonstrate that while there is considerable heterogeneity in the details of context dependence for individual cells in both cortex and thalamus, on average the effects are similar across subdivisions of the thalamus. In contrast, nonlinear context dependence of auditory cortical responses differs between A1 and AAF, with greater simultaneous enhancement in A1 and faster delayed suppression in AAF.

4.2 Materials and Methods

These experimental methods were similar to those described by Linden et al. (2003).

4.2.1 Animals

Twelve adult CBA/Ca mice (6-15 weeks old) were used to gather cortical data, and six adult CBA/Ca mice (6-8 weeks old) were used to gather the thalamic data.

4.2.2 Surgical Procedures

Cortical surgical procedures conformed to protocols approved by the University of California at San Francisco's Committee on Animal Research and were in accordance with federal guidelines for care and use of animals in research. Thalamic surgical procedures were similar and were performed in accordance with the United Kingdom Animal (Scientific Procedures) Act of 1986.

Mice were anaesthetised and maintained at a surgical plane of anesthesia through the use of ketamine and medetomodine. An initial intraperitoneal bolus injection of anesthetic was given to sedate the animal. Following this, a canula was placed into the animal's peritoneum so that further boluses or continuous infusion of anaesthetic could be provided. Dexamethasone was administered to control brain oedema, atropine to minimise bronchial secretions, and Ringer solution to ensure adequate hydration. The animal was kept on a homeothermic blanket (Harvard Instruments) to ensure that the body temperature was maintained at approximately 37.5°C (monitored via a rectal probe). Once fully anaesthetised and prepared for surgery, the animal was placed onto a bite bar in order to immobilise its head after which the skin was transected along the midline to expose the skull.

For cortical recordings, a small craniotomy was performed on the left-hand side of the skull, above the known location of auditory cortex (bordered rostrally by the lambdoid suture, caudally and ventro-laterally by the squamosal suture, and dorso-medially by the temporal ridge). For thalamic experiments, a craniotomy approximately 2.5mm in diameter, centred 2.75mm lateral to midline and 2.75mm caudal to bregma, was performed on the right-hand side of the skull, enabling vertical access to the thalamus. In both cases, the cortical surface was kept moist by the regular application of warmed saline.

4.2.3 Recording Procedures

All experiments, cortical and thalamic, were conducted in a sound-shielded anechoic chamber (Industrial Acoustics).

Auditory stimuli were directed towards the animal's ear contralateral to the craniotomy via a free-field speaker, and a sound-attenuating plug was placed in the ipsilateral ear. Prior to the start of each experiment, acoustic stimuli were calibrated with a Bruel and Kjaer microphone positioned near the opening of the animal's auditory canal. Typically, this ensured that the sound system's frequency response was flat to within ± 1dB from 2-90 kHz.

For cortical experiments, extracellular recordings were made using epoxylite-coated tungsten electrodes (1-4 MΩ impedance). These were introduced into the left auditory cortex in penetrations orthogonal to the cortical surface. Recordings targeted the thalamorecipient layers III/IV (Smith and Populin, 2001) by cortical depth (350-600 μm below the dural surface). Cortical areas were found and identified as described by Linden et al. (2003).

For thalamic experiments, extracellular recordings were made across all thalamic subdivisions using custom-made linear arrays consisting of eight WPI tungsten electrodes (impedance typically 1-2 MΩ). The array was placed perpendicular to the midline with the first penetration targeting a position approximately 2 mm from midline and 3 mm from bregma, as this position was deemed most likely to yield responses from all three major thalamic subdivisions (Anderson and Linden, 2011). The electrode was first moved down to 2200μm below the cortical surface, and then left to stabilise (to allow electrode induced brain movement to cease) for ~10 min. Neurons responsive to auditory stimuli were located through the use of a 50 μs click presented at ~60 dB SPL. Once an auditory response had been established (typically at 2900 μm), further sites were located by progressing the electrode 100 μm at a time, until auditory activity was lost.

4.2.4 Histological Procedures

Histological delineation was carried out for all thalamic recordings. Procedures were similar to those described by Anderson et al. (2009a).

Electrolytic lesions were created by passing current through the desired electrode on the array (5μA for 7 secs). Such lesions were typically created at the most medial and lateral electrodes on the array that yielded auditory activity. This was replicated at both the top and bottom of the electrode track. Ideally, this procedure yielded four lesions (two at the top of the track, and two at the bottom), bracketing the area over which auditory activity was located. This allowed for estimation of shrinkage and histological reconstruction of most recording sites.

Once lesioning had taken place, animals were given an overdose of barbiturate anaesthesia (sodium pentobarbital) and perfused transcardially with 4% paraformaldehyde in 0.1 M phosphate buffer. Following perfusion, the brain was removed and placed in the paraformaldehyde solution for 1-2 days. Blocks containing the full auditory thalamus were then cut into 50μm slices using a vibrotome. The sections were then stained for the metabolic marker, cytochrome oxidase (CYO). To demonstrate expression, slides were incubated for 3-7 hours at 37° in a solution containing 20 mg of diaminobenzidine hydrochloride in 10 ml of distilled water and 30 mg of cytochrome c with 3 g of sucrose in 30 ml of 0.1 M phosphate buffer.

Electrolytic lesions were visualised in the stained brain sections using a Zeiss AxioPlan 2 Imaging microscope (magnification x25-x200). The position of each neuron was assigned to the appropriate subdivision as defined by the CYO distribution. Ambiguous recording sites were not included in the subdivided data.

4.2.5 Stimuli

4.2.5.1 Simple Stimuli

Simple tonal stimuli consisting of 50 ms tone pulses, ramped up and down with 5 ms cosine gates, were used to characterise the frequency response area (FRA) of the neural sites. The frequency and intensity of each tone were varied pseudorandomly over the range of possible values in the stimulus set. In cortical experiments, frequencies spanned either the range 2-32 kHz (low frequency stimulus set), or the range 25-100 kHz (high frequency stimulus set). In thalamic experiments, only the low frequency stimulus set was used. Intensities ranged from 0-70 dB SPL in 5 dB increments. Each of the possible frequency-intensity combinations was presented only once per stimulus set.

In addition to the use of tonal stimuli to characterise frequency-intensity sensitivities, a selection of other simple stimuli, including clicks, broadband noise, and frequency-modulated sweeps, were utilised to identify sites where auditory activity was present.

4.2.5.2 Complex Stimuli

For both cortical and thalamic experiments, dynamic random chord (DRC) stimuli (described previously by Linden et al. (2003)) was utilised. This complex stimulus consists of a series of spectrotemporally-rich random chords. The stimulus is clocked, such that every 20 ms, a combination of 20 ms cosine-gated tone pulses with randomly chosen frequencies is generated. The centre frequencies of the tone pulses were chosen from 24 or 48 different possibilities (25-100 kHz or 2-32 kHz, respectively; cortical data) or 48 different possibilities (2-32 kHz; thalamic data). The number of tones that made up a chord was random, with an average spectrotemporal density of two tone pulses per octave. The peak level of each pulse was chosen randomly from 10 different intensity levels, 5 dB-SPL apart in the range 25-70 dB-SPL. A single trial of such a stimulus lasted 60 seconds. Full presentation of the stimulus lasted for 20 minutes, allowing for 20 continuous trials.

Spike times collected during presentation of such stimuli were analyzed off-line using Bayesian spike-sorting techniques (Sahani, 1999; Lewicki, 1998), to extract responses from either small clusters of neurons (for the most part) or, occasionally, single-units.

4.2.6 Modelling Neural Responses to Sound

Much of the modelling that is utilised was discussed at length in chapter 3.

Briefly, we fit both linear and multilinear models to the DRC-evoked neural responses. The STRF model was discussed in section 2.2.2, and the stimulus-response function (the function relating the stimulus spectrogram to the neural response) was given by equation (2.1). Estimation of the STRFs was carried out using the automatic smoothness determination algorithm (ASD) algorithm (Sahani and Linden, 2003a). This technique was discussed in section 3.4.4. Conceptually, this estimation procedure amounts to estimating the optimal amount of spectrotemporal smoothing to apply to the STRF during a regularised linear regression.

The mathematical details of the multilinear framework that we utilise here was the focus of chapter 3. We use the models described by equations 3.24 and 3.27. Estimation was carried out using the alternating least squares (ALS) procedure, discussed in section 3.4.1. These models will be treated in more detail later in this chapter.

4.2.7 Predictive Capability of Neural Encoding Models

In order to evaluate the predictive power of a neural encoding model, a standard approach is to use some measure of explainable variance; that is, some statistic that tells us how much variability within the observed signal we are able to capture with our model prediction. A standard statistic for such a purpose is the coefficient of determination (or R^2), given by $R^2 = (P(\text{total}) - P(\text{error}))/P(\text{total})$, where P is used here to denote power (variance over time). $P(\text{total})$ refers to the power in the observed signal, and $P(\text{error})$ refers to the power in the error (or residual). Normalising this difference by the total power yields a value between 0 and 1, where 1 dictates that all of the variance has been captured.

Neural data are noisy and perfect prediction of a noisy signal is, by definition, impossible. Thus, a statistic such as R^2 is ill-suited for neural data. Here, we exploit the fact that we are using multi-trial data, and utilise the signal power statistic (Sahani and Linden, 2003b), which provides an estimate of the stimulus-related variability within the observed signal (the component of the signal which we should, in theory, be able to predict). We use this statistic as an alternative denominator in a pseudo-R^2 statistic; this, we will refer to as predictive power.

4.2.8 Neuronal Populations

We used the signal power metric of Sahani and Linden (2003b) to establish which of our neuronal recordings exhibited a significant amount of stimulus-related variability and were worth utilising for further analysis. We discarded all recordings that did not have a signal power at least 1 standard deviation away from zero.

This left us with populations of neuronal responses to dynamic random chord stimuli recorded in 82 cortical sites and 122 thalamic sites. The cortical sites can be further subdivided into 39 sites located in A1, and 43 in AAF. The thalamic sites can be further subdivided into 11 from the dorsal subdivision, 34 from the medial subdivision, and 51 from the ventral subdivision. Note here that not all 122 thalamic sites were able to be attributed to a particular subdivision, due to histological ambiguity.

4.3 RESULTS

4.3.1 MODELLING NEURAL RESPONSES IN THE AUDITORY CORTEX AND THALAMUS

For decades, the STRF model has been used as a standard tool for modelling the stimulus-response function of neurons within the auditory system (e.g., Aertsen et al. (1981); Aertsen and Johannesma (1980); Aertsen et al. (1980); deCharms et al. (1998); Fritz et al. (2007); Depireux et al. (2001); Woolley et al. (2005)). The STRF represents a linear estimate of a neuron's selectivity for the spectrotemporal features of a sound stimulus. If the true stimulus-response function of the neuron is nonlinear, then such a linear estimate will be inherently stimulus-dependent (Theunissen et al., 2000; Christianson et al., 2008). This flavour of stimulus-response analysis has its roots in the classic Volterra/Wiener series expansion (Volterra, 1930; Wiener, 1958). Briefly, such a Volterra series expansion (in the discrete setting) provides an estimate of the time varying firing rate $\mathbf{r}$ as

$$\mathbf{r} = k_0 + \mathbf{k}_1.\mathbf{x} + \mathbf{x}^T\mathbf{K}_2\mathbf{x} + \cdots \tag{4.1}$$

An STRF (or Wiener filter, given white noise input) provides an estimate of the first order kernel $\mathbf{k}_1$. Higher order kernels can also readily be estimated using linear regression (with inputs augmented to reflect the kernel order; a second-order kernel would require all quadratic combinations of the input, for example). Typically, gathering adequate data to go beyond second-order is difficult however, due to the dramatic increase in parameter count. As a result of this, an ideal solution could be to define a second order model, but to restrict the second order interactions in some way, such that it can become easier to estimate using limited data.

Here, we focus on a extended version of the context model, originally described by Ahrens et al. (2008a). This model has been shown to be capable of capturing nonlinear stimulus interactions such as combination sensitivity and forward suppression. Previously, the model included fully separable contextual interactions, and two input nonlinearities. Here, we remove both input nonlinearities, and allow both receptive field components within the model to be inseparable. The stimulus-response function of the

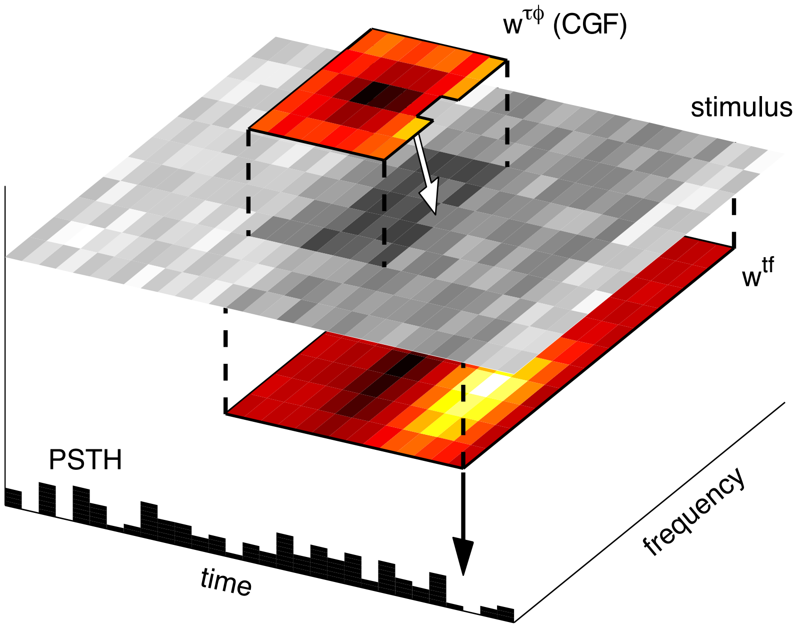


Figure 4.1: Extended context model. The diagram shows a contextual gain field ($\mathbf{w}^{\boldsymbol{\tau\phi}}$), a stimulus (dynamic random chord, discretised in time and frequency), and a primary receptive field ($\mathbf{w}^{\mathbf{tf}}$; an STRF-like field). The CGF acts to multiplicatively modulate the effective sound level of each target tone within the stimulus (*blue arrow*) before the primary receptive field linearly transforms the effective sound levels (*green arrow*) to an estimate of the firing rate.

model is given by

$$\hat{r}(i) = c_1 + \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk}^{\mathbf{tf}} s(i-j+1,k)\left(1 + \sum_{m=0}^{M}\sum_{n=-N}^{N} w_{mn}^{\boldsymbol{\tau\phi}} s(i-j+1-m,k+n)\right) \quad (4.2)$$

This equation yields a prediction of a neural firing rate $\hat{r}$ at some time i. The model consists of a linear component, with a principal receptive field (PRF; analogous to an STRF) denoted by $\mathbf{w}^{\mathbf{tf}}$, and a contextual gain field (CGF) denoted by $\mathbf{w}^{\boldsymbol{\tau\phi}}$, which acts to multiplicatively modulate the stimulus spectrogram prior to spectrotemporal summation by the PRF (shown schematically in figure 4.1).

The superscripts (in bold) of the receptive field components correspond to their physical dimensions, with the italicised subscripts denoting their corresponding index. The superscripts $\mathbf{t}$ and $\mathbf{f}$ correspond to the PRF dimensions of time-lag and frequency, and are indexed by j and k, respectively. The superscripts $\boldsymbol{\tau}$ and $\boldsymbol{\phi}$ correspond to the CGF dimensions of *relative* time-lag and *relative* frequency (with respect to a given tone within the stimulus), and are indexed by m and n. The upper limits of the j and k summations simply denote the maximum dimensions of the PRF (the maximum time-lag (J), and frequency extent (K) of the receptive field). The summations over m and n are subtly different. The m summation begins at 0 (rather than 1), in order to incorporate a contextual time-lag of 0 into the model (thus the CGF $\boldsymbol{\tau}$ dimension will be $(M+1)$). The subscript n corresponds to frequency deviation and thus can be either positive or negative, with N denoting the maximum allowed deviation, leading to a CGF $\boldsymbol{\phi}$ dimension of $(2 \times N + 1)$.

The intuition behind the model is that the CGF essentially defines an acoustic neighbourhood (a local context) around each tone within the stimulus spectrogram. The weighting of this neighbourhood is then used to multiplicatively modulate the intensity of the given tone. This operation is carried out for every tone within the stimulus, before a linear (STRF-like) prediction is generated. Thus, the predicted response of the neuron at time i will be influenced by the local acoustic context present within the stimulus.

Being linear in only first and second order multiplicative interactions, such a context model has to be similar to a second order Volterra model, as given by equation (4.1). The key difference however, is in the parametrisation of the context model. The parametrisation that we use imposes specific structural limitations on the range of possible second-order interactions. These structural limitations are designed in such a way that they mimic the nonlinear effect of stimulus context.

4.3.2 Uniformity of Contextual Interactions

We previously discussed how our goal was to work with a model framework, similar in flavour to a full second-order Volterra model, but where we have structurally limited the range of possible second-order interactions to allow for easier estimation given limited data.

A specific assumption of the context model as we have defined it in equation (4.2) is that the context field is completely invariant with respect to different time-frequency positions within the PRF. Specifically, we assume that contextual effects are equal at all frequencies and time lags, i.e. a single CGF operates over the entire domain of the PRF. This is not necessarily a valid assumption to make however, and we wish to directly test this hypothesis of contextual uniformity before proceeding to further analysis.

In order to achieve this, we extended the context model further, such that multiple CGFs can be used, and that each has the scope to act on a different subset of the PRF. In theory, this allows for contextual effects to be different for different combinations of frequency and time-lag within the PRF.

Equation (4.2) is augmented as follows

$$
\begin{aligned}
\hat{r}(i) \quad = \quad & c_1 + \\
& \sum_{\{j,k\}\in\rho_1} w^{\mathbf{tf}}_{jk} s(i-j+1,k) \left(1 + \sum_{mn} w_{\mathbf{1}}{}^{\boldsymbol{\tau\phi}}_{mn} s(i-j+1-m,k+n)\right) + \\
& \sum_{\{j,k\}\in\rho_2} w^{\mathbf{tf}}_{jk} s(i-j+1,k) \left(1 + \sum_{mn} w_{\mathbf{2}}{}^{\boldsymbol{\tau\phi}}_{mn} s(i-j+1-m,k+n)\right) + \\
& \quad \vdots \\
& \sum_{\{j,k\}\in\rho_s} w^{\mathbf{tf}}_{jk} s(i-j+1,k) \left(1 + \sum_{mn} w_{\mathbf{S}}{}^{\boldsymbol{\tau\phi}}_{mn} s(i-j+1-m,k+n)\right)
\end{aligned}
\tag{4.3}
$$

where $\rho_1 \cdots \rho_s$ are sets that contain different time-frequency ($\{j,k\}$) pairs of interest.

Here, the original stimulus-response function has been split into a sum of S terms, where S dictates the number of CGFs present within the model. This equation has some particularly interesting parallels to the Volterra approach that has been previously discussed. Specifically, if S is equivalent to the actual number of $\{J,K\}$ elements within the PRF (each time/frequency element has its own CGF associated with it), then such a split context model is equivalent to a second order Volterra model, in that all second order interactions will be captured. Of course, the number of parameters involved in a such an estimation is huge (we have $J \times K$ time-frequency elements, and each of these has an associated $M \times N$ CGF). With a such a large parameter count, we also need to be particularly careful when evaluating the predictive capabilities of such a model, since estimating such a large number of parameters leads to the problem of overfitting (which essentially amounts to the "explanation" of noise). As a result of this, we chose to examine a number of simpler split models, where we segregated the PRF into two divisions, and allowed a CGF to be associated with each of them. Even though such a split is a clearly a large distance away (in modelling terms) from associating a separate CGF with every time-frequency element, this model still allows us to test specific hypotheses about whether the CGFs might differ when associated with different parts of the PRF. Importantly however, the parameter count is such that it becomes far more manageable.

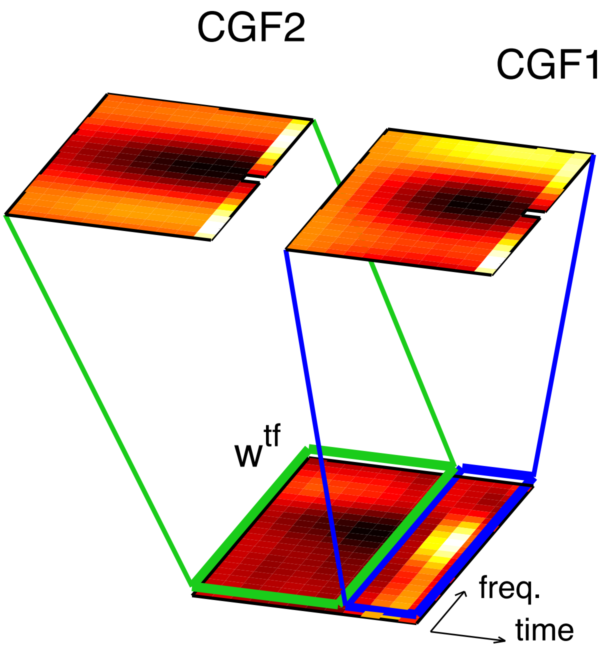


Figure 4.2: Split context model. This diagram illustrates schematically, the model given by equation (4.3). Here, the model consists of two CGFs ($\mathbf{w_1^{\tau\phi}}$ and $\mathbf{w_2^{\tau\phi}}$), and a PRF ($\mathbf{w^{tf}}$). Each CGF acts on a different subset of the underlying PRF (in this illustration, the subsets correspond to the presence of excitation and inhibition). Although only two CGFs are shown here for simplicity, in principle one can specify a model with a far larger number (although this will increase the model parameter count dramatically, and illustrates the need for large amounts of data to prevent overfitting).

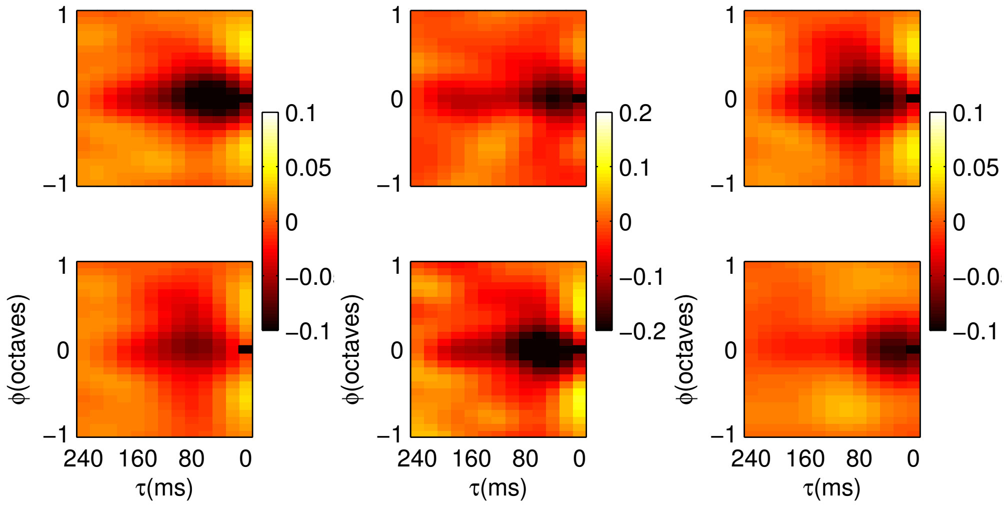
In testing for contextual uniformity in such a way, there are a number of different results that one could observe. Ultimately, the most obvious of these is to do with the parameters themselves. If, for example, a model is fit with (say) twenty CGFs, and it turns out that every CGF contains similar parameters, then this would serve to suggest that the model was more complex than needed, and a simpler representation would be the correct one. A similar argument could also be made with predictive power in that, if one adds additional parameters to a statistical model, if those parameters are actually useful, then the cross-validated predictive power should increase. There are, of course, some potential issues here in regards to overfitting (some of these issues, and how they can be handled were discussed in section 3.4.4).

4.3.2.1 Structural Similarity

We fit a variety of different two-CGF context models to both our cortical and thalamic populations of data. For the vast majority of the models that we fit, we noticed a remarkable similarity in the structure present between pairs of CGFs. To illustrate this, figure 4.3 (a) and (b) show cortical population averages of two two-CGF models. The first of these (in (a)) is a model where one CGF (top row) has been applied to purely the excitatory portion of the underlying PRF, while the second CGF acts upon the entire range of the PRF. In this example, aside from a change in magnitude, the structure seems to be qualitatively similar, with this population average exhibiting a large delayed suppressive region, and two noticeable regions of near-simultaneous ($\tau = 0$) enhancement. This qualitative similarity also seems to persist in another split that was tried, shown in (b). Here, the top CGF was applied only to the low-frequency half of the PRF, and the bottom CGF only to the high-frequency half. Again, both a delayed suppressive region, and a region of near-simultaneous enhancement can be observed.

We were particularly interested however, in directly testing whether contextual effects that act upon the excitatory and inhibitory components of the PRF are similar. This was one of the only splits that we tried in which we actually seemed to observe a qualitative difference in the population average structure. This is illustrated in figure 4.3 (c). Here, there are certainly some aspects of similarity. There is a delayed suppressive region present in both CGFs for example, although the timescale is somewhat different in the CGF fit to the inhibitory part of the PRF (bottom row). The near-simultaneous enhancement clearly observable in the excitatory CGF (top row) is also not clearly present within the inhibitory CGF.

To probe these differences further we carried out a detailed cell-by-cell analysis of the excitatory/inhibitory split model, in order to further test our hypothesis of contextual uniformity. Even though, on average, some differences in structure were observed, we were interested in assessing whether or not the similarity between two CGFs belonging to a single cell was comparable to the similarity between two random CGFs drawn from the population. In order to quantify similarity, an uncentered correlation coefficient was used, which yields the cosine of the angle between the two CGFs. A value of 1 indicates perfect correlation, whilst a value of -1 indicates perfect anti-correlation. Figure 4.4 (a) shows the distributions of these correlation coefficients for random pairs of CGFs within the population (in dotted black) and the correlation coefficients between pairs of CGFs for the same cell (in black). The background distribution here is largely

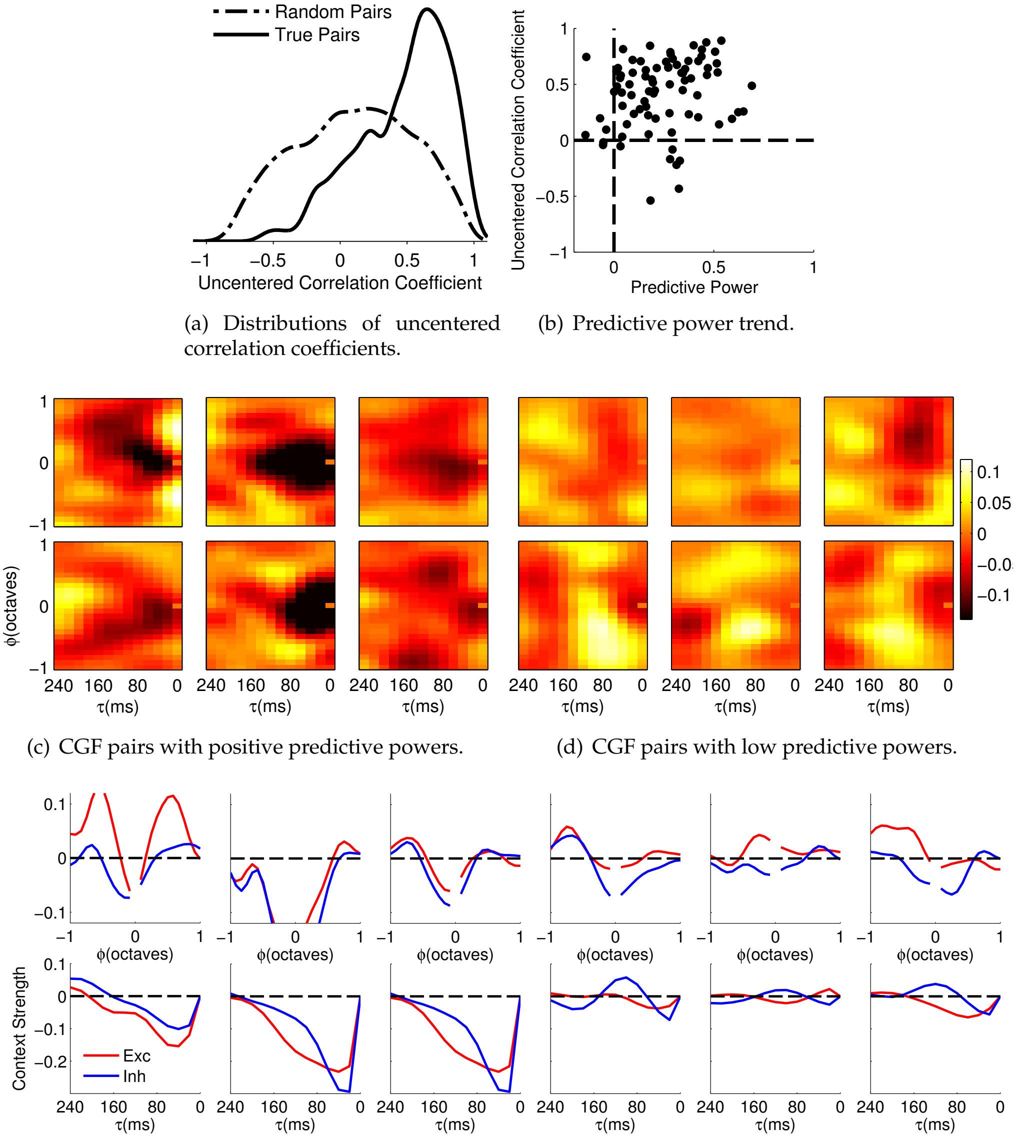


(a) Excitatory/Background split. (b) Low/high frequency split. (c) Excitatory/inhibitory split.

Figure 4.3: Population two-CGF model fits. This figure illustrates the qualitative similarity between the population averages of some two-CGF model fits. (a): A model where CGF1 (*top*) acts upon the excitatory portion of the PRF and CGF2 (*bottom*) acts on the entire PRF. (b): A model where CGF1 (*top*) acts upon the low-frequency half of the PRF and CGF2 (*bottom*) acts upon the high-frequency half. Notice that in both (a) and (b), the structure between CGFs is qualitatively similar (up to a change in magnitude, noticeable due to the same scale being used on both pairs). (c): A model where CGF1 (*top*) acts on the excitatory part of the PRF, and CGF2 (*bottom*) acts on the inhibitory part. These population averages share a similar suppressive region (with a temporal difference), but the enhancement present at $\tau = 0$ in CGF1 is not present in CGF2.

concentrated around 0, indicating that if a random pair of CGFs from the population is chosen, they are most likely to have very little similarity to one another. Conversely, the true distribution, representing CGF pairs from the same cell, is clearly skewed in a positive direction away from 0. This quite clearly indicates that CGF pairs from the same cell are likely to be highly correlated to one another.

Figure 4.4 (b) - (e) demonstrate that this similarity depends heavily upon the predictive capability of the model fit. In (b) we show the predictive power of the model plotted against the correlation coefficient. Although outliers do exist, the general trend within the data is that as the predictive power increases, so does the degree of similarity between CGFs. In (c), we have included the CGF pairs for three model fits yielding a high predictive power (0.69, 0.40, 0.47, from left-right). In (d), the fits yielded far lower predictive powers (0.0005, 0.0199, 0.0153, from left-right). What should be obvious from the representative pairs in (c) and (d), is that clear structure seems to exist in the CGFs of models that can predict well, and no discernible structure within the CGFs in mod-



(a) Distributions of uncentered correlation coefficients.

(b) Predictive power trend.

(c) CGF pairs with positive predictive powers.

(d) CGF pairs with low predictive powers.

(e) Temporal/spectral slices of the above CGF pairs.

Figure 4.4: Excitatory/inhibitory split model - single cell examples. (a): Distributions of correlation coefficients indicating the level of similarity between two CGFs chosen at random from the population (in dotted black), and the level of similarity between two CGFs belonging to a single cell (in black). Clearly, the black distribution is skewed in the positive direction indicating that CGF pairs from the same cells are likely to be highly correlated, more so than a random pair from the population. (b): CGFs are more likely to be similar if the predictive power is high. (c): Three CGF pairs for model fits yielding high predictive powers (0.69, 0.40, 0.47, from left-right). (d): Three CGF pairs for model fits yielding low predictive powers (0.0005, 0.0199, 0.0153, from left-right). (e): Characteristic structure extracted from the CGFs located above. *Top row*: A spectral strip at $\tau = 0$, for all ϕ, the near-simultaneous region corresponding to a time-lag of 0-20 ms. *Bottom row*: A temporal strip at $\phi = 0$, for all τ. Notice that these features are remarkably consistent within the high predictive power pairs, but significantly less so in the negative predictive power pairs.

els that cannot predict well. Figure 4.4 (e) clarifies this further by showing some of the characteristic structure from the corresponding CGFs in the rows above. The top row details the structure present at $\tau = 0$, the vertical strip corresponding to a near-simultaneous time-lag of 0-20 ms, at all relative frequencies. The bottom row shows a temporal strip, centered at $\phi = 0$, and extending through all values of τ. It should be relatively clear that these particular features are remarkably consistent within the high predictive power CGF pairs. Consistent structure in the negative predictive power pairs is far less obvious.

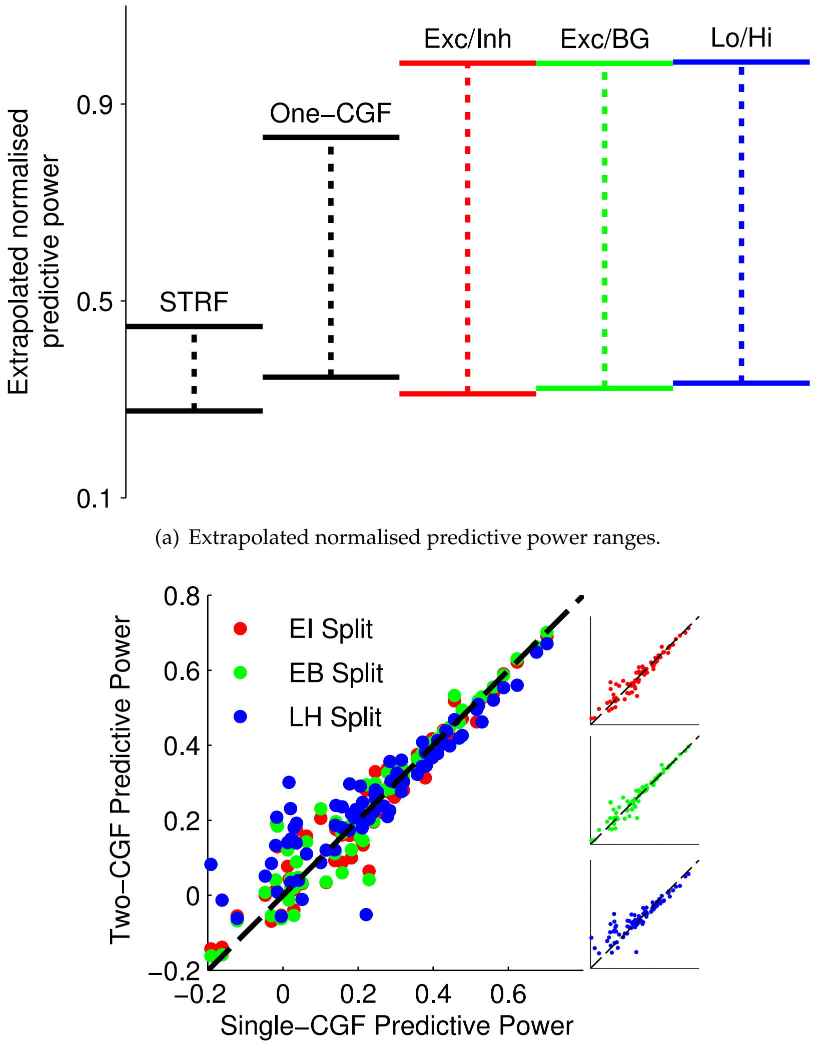
4.3.2.2 Predictive Capability of Multi-CGF Context Models

Having established a certain amount of similarity in the structure of the CGF pairs, we were curious to establish how predictive these models were, especially when compared to a single CGF model, with its assumption of contextual uniformity.

To investigate this, we replicated the analysis of Sahani and Linden (2003b) (also Ahrens et al. (2008a)) and extrapolated both cross-validated (test-set) predictive powers and training-set predictive powers to zero noise power (where the noise power is computed similarly to the signal power estimator discussed earlier, and provides an estimate of the temporal variability due to noise). These predictive powers can be treated as lower and upper bounds on what can possibly be achieved by each model (for details, see Sahani and Linden (2003b)). Figure 4.5 (a) shows schematically the different predictive power bounds achieved by fitting a standard linear STRF model, a single-CGF context model, and the three two-CGF context models presented earlier, to a population of cells within the auditory cortex.

From the perspective of overfitting, as the number of parameters increases in a statistical model, the predictive power on the training set alone should increase (due to a reduction in training error attributable to the parameter increase). This is certainly the case here, as the extrapolated training-set predictive power is at its lowest for an STRF model, and gradually increases as one, and then two, CGFs are added. Perhaps unsurprisingly, the extrapolated training-set predictive powers are almost identical for all three two-CGF models. Even though the models are specified differently, they all contain an identical number of parameters.

Conversely, cross-validated predictive power should only increase with the number of model parameters, if the parameters are actually useful (since cross-validation can be thought of as a measure of how well the model is capable of generalising). Thus,



(a) Extrapolated normalised predictive power ranges.

(b) Predictive power comparison for three split models.

Figure 4.5: Predictive capability of multi-CGF models. (a): Extrapolated predictive power ranges. The upper bounds here are provided by extrapolation of the training-set predictive powers, whilst the lower bounds are provided by extrapolation of the cross-validated (test-set) predictive powers. The upper bounds increase as a function of parameter count, as is expected. The lower bounds for the four context models are greater than that of an STRF model, indicating that they all provide an increase in quality over a simple linear fit. The lower bounds of all three multiple-CGF models are lower than that of the single-CGF model, further emphasising the point that an additional CGF does not seem to provide an increase in model quality. (b): A direct comparison of the cross-validated predictive powers achieved through fitting a single-CGF context model, or a two-CGF context model. Colours denote the different splits. *Red*: Excitatory/inhibitory. *Green*: Excitatory/background. *Blue*: Low frequency/high frequency. For a large number of cells within the population, the predictive power lies either on or below the $y = x$ line, indicating that the increase in complexity provided by an additional CGF does not improve the quality of the model fit.

both single, and two, CGF context models provide a (perhaps subtle) increase over a simple STRF model. Of particular interest however, is the fact that the extrapolated cross-validated predictive power actually goes down very slightly when a second CGF is added (in all three cases). This indicates that the significant increase in complexity does not add much to the quality of the overall model fit. This is probably due, in large part, to the amount of similarity between both CGFs in this split formulation of the context model. Figure 4.5 (b) serves to emphasise this point by directly plotting the cross-validated predictive powers of the different context models against one another. For all three split models, a large number of the points lie either on or beneath the $y = x$ dotted line, indicating that, for most cells, the addition of a second CGF does not provide an increase in predictive capability.

Ultimately, these analyses have all been carried out in order to determine whether or not the assumption of contextual uniformity within the single CGF context model is a valid one. Can we use a model where we have the assumption that contextual interactions are identical over multiple frequencies and time-lags? For the most part, this certainly seems to be the case, and this particular structural limitation on second-order interactions is a valid modelling assumption to make. As a result of this, all of the following analyses will focus on the single-CGF model, as denoted by equation (4.2).

4.3.3 Contextual Gain Fields in Cortex and Thalamus

We proceeded to fit single-CGF context models to populations of data recorded from both the auditory cortex and thalamus.

Aside from exclusion of the input nonlinearity part of the model, the crucial extension in this version of the context model is that we allow receptive field components to be inseparable. As a result of this, we were particularly interested in carrying out a detailed structural analysis on the estimated CGFs, since they can, in principle, provide us with significant insight into how combination sensitivities manifest themselves within the stimulus evoked neural response.

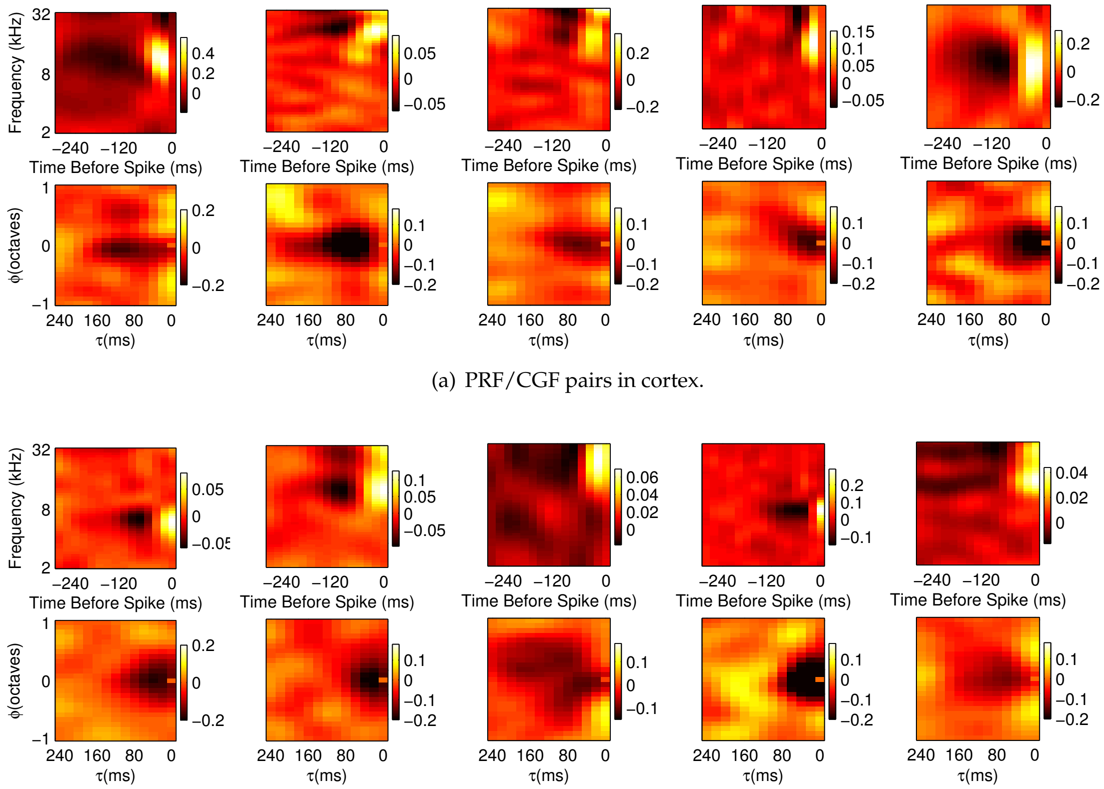
4.3.3.1 Model Interpretation

Interpretation of any structure present within a CGF is particularly important if one wishes to make claims about underlying biological mechanisms. In the previous section we detailed why the assumption of contextual uniformity present within the single-CGF model was a valid assumption. Here, we will revisit this point, as it is important for an accurate interpretation of model structure. The values within a CGF are best thought of as providing a *relative gain*, since the modulated stimulus values still have to be linearly filtered, and the resultant prediction depends on the sign of the underlying linear filter. For example, suppose a CGF has a positive value at $\tau = 0$, some half octave in frequency above the current tone of interest. This means that if we have a tone in our stimulus, and a second tone is played simultaneously, a half-octave above, then the weighting of this tone in the convolution with the spectrogram should be increased. This will then result in a facilitatory effect on predicted firing rate if the underlying PRF contains a positive weight, or a suppressive effect on predicted firing rate if the underlying PRF contains a negative weight. Basically, positive values will be made more positive, and negative values will be made more negative. Of course, the opposite is true if the value within the CGF is negative; this would lead to a decrease in excitation, or a decrease in inhibition, depending on the sign of the PRF. Thus, the positive and negative weights in the CGF indicate enhancement or suppression of the linearly filtered response, not excitation or inhibition.

4.3.3.2 Structural Aspects of CGFs

One of the fundamental results that we would like to present is that the CGF structure, observed in both cortex and thalamus, indicates that contextual interactions are heavily dependent on the precise relative spectrotemporal arrangement of sound energy within a complex stimulus. Specifically, both near-simultaneous and delayed sound energy seems to play an important role in sound processing in both the auditory cortex and thalamus. This structure will be analysed in greater detail throughout the rest of this chapter.

We initially carried out a cell-by-cell analysis of the single-CGF model fits. Several examples of this are provided in figure 4.6. Figure 4.6 (a) shows five PRF/CGF pairs from the cortical population. As is to be expected, the PRF component of the model closely resembles STRF structure previously reported in the mouse auditory cortex (Linden et al., 2003). PRF and STRF properties will be directly compared later in this chapter. One of the most noticeable aspects of the CGF examples in (a) (and also in (b), from the thalamic population), is that similar structure appears consistently throughout both populations. This is the same structure that was observed in the split model fits earlier, and consists of a delayed suppressive region and some form of near-simultaneous enhancement at $\tau = 0$.



(a) PRF/CGF pairs in cortex.

(b) PRF/CGF pairs in thalamus.

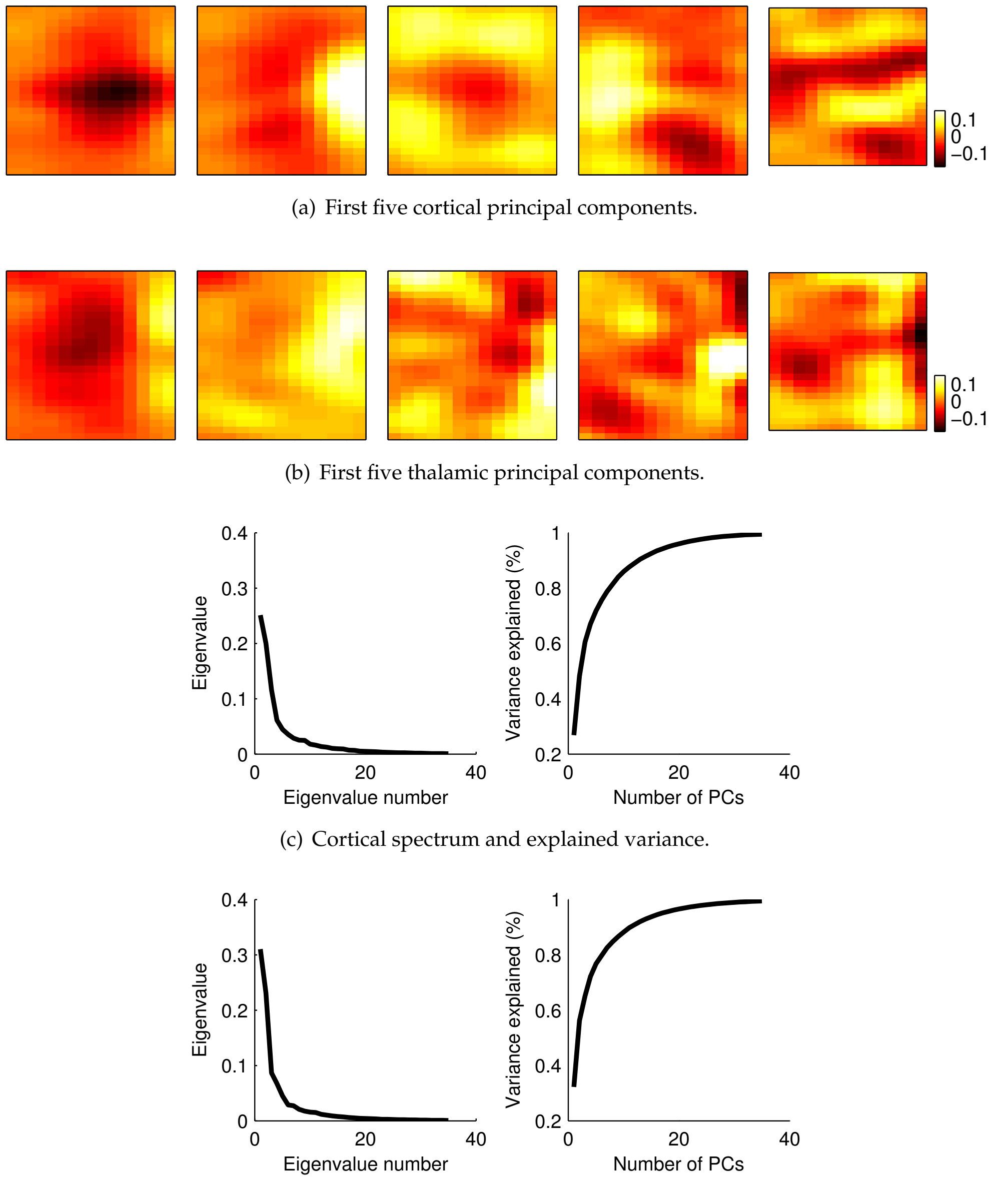
Figure 4.6: CGF single cell examples. (a) and (b): Five examples chosen from the cortical and thalamic populations. Notice that the inhibitory subfield widths tend to be shorter in thalamus than in cortex. For the most part, the structure present within the CGFs, even at the single cell level, is remarkably consistent across the population. In the cortical examples presented in (a), almost all of the CGFs show some form of delayed suppression and near-simultaneous enhancement. This is typically similar in the thalamic examples (b), although the structure seems less consistent.

An important point is that these two structural features are by no means the only significant structure that appears within individual CGFs (as can be quite clearly observed in both (a) and (b)). Our primary reasoning for focussing on the delayed suppression and the near-simultaneous enhancement, is purely due to its remarkable consistency across different cells. This will also become particularly evident when population CGF structure is presented, later in this section. It is certainly the case that perhaps all of the structure evident within each individual CGF could be relevant, and indicative of the individual spectrotemporal response properties of particular cells. This is at least somewhat consistent with the ideas of Sadagopan and Wang (2009), who show that neurons within the marmoset auditory cortex can be particularly sensitive to the precise spectrotemporal combination of tone pips.

4.3.3.3 Assessing CGF Population Variability

In order to get some indication of how variable this structure was, we applied a principal components analysis (PCA) to all CGFs within both populations. Briefly, a data matrix was constructed for each population, where each row within the data matrices represents an individual CGF. We then centered the data prior to calculating its covariance matrix. The principal component analysis itself amounts to performing an eigendecomposition on the covariance matrix, in order to yield a set of eigenvectors that are representative of the directions in which the CGFs differ from the mean CGF. The first five of these directions, for each population, are shown in figure 4.7 (a) and (b). The cumulative amount of variance explained can be simply calculated as the cumulative sum of the eigenvalue spectrum, normalised by its sum. This is shown in 4.7 (c) and (d). In both cortex and thalamus, the first five principal components (as shown) are responsible for explaining 72% and 77% of the variance, respectively.

The structure present within these principal components is of particular importance in understanding where, within a CGF, variability is likely to occur. The first principal component in the cortical population actually shows structure similar to what we have observed in the population CGF examples thus far (in figure 4.3). This implies that it is this structure that is also the most variable over the population. This does actually correspond somewhat, to the differences we observe when analysing the CGFs of single cells. Although there is consistent structure over multiple cells, one of the most noticeable differences is a large diversity in the magnitude of the CGF weights (see figure 4.6 for specific examples of this). In the thalamus, the results are somewhat similar, in that the first principal component indicates that the greatest variability occurs within the delayed suppressive and near-simultaneous regions of the CGF. Interestingly though, the first few principal components actually account for more variability within thalamus than they do in cortex, suggesting that the variability across the population may be more constrained (dimensionally at least) in thalamus than in cortex. This could perhaps be indicative of the level of complexity present within cortical, as opposed to thalamic, responses. Such an increase in accountable variability might be explained by individual cells within the cortical population having more varied spectrotemporal preferences than observed in the thalamic population.



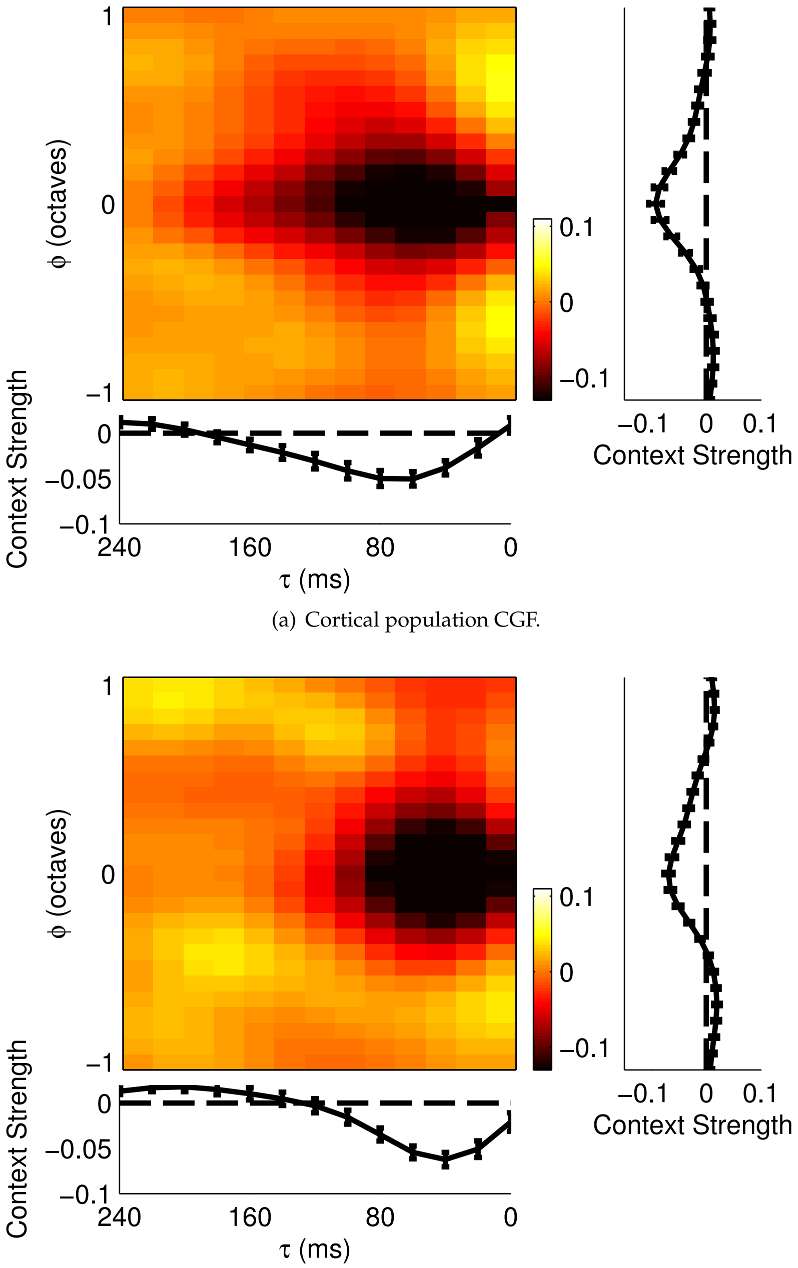
(a) First five cortical principal components.

(b) First five thalamic principal components.

(c) Cortical spectrum and explained variance.

(d) Thalamic spectrum and explained variance.

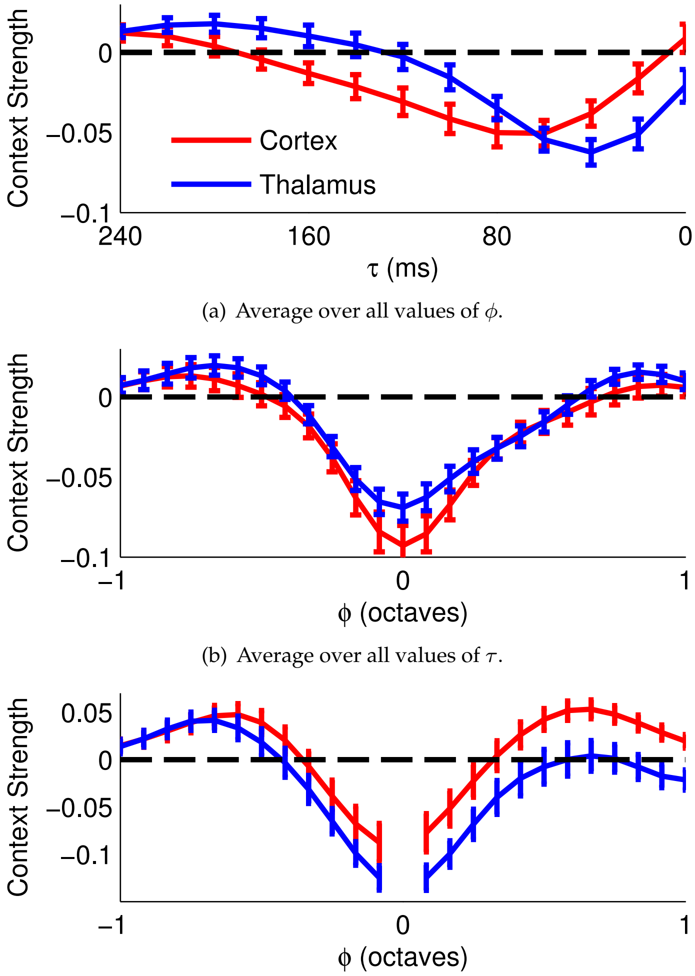
Figure 4.7: CGF population variability. Variability across the populations was quantified using principal component analysis. This amounts to performing an eigendeomposition on the covariance matrix of the data (where each row in the data matrix is a separate CGF). The resultant eigenvectors (the principal components) can be interpreted as directions in which the CGFs vary from the mean CGF. (a) and (b): The eigenvectors corresponding to the five largest eigenvalues from each spectrum. The spectrum itself, along with the amount of variance explained (cumulatively) via the addition of each principal component, is shown in (c) and (d). In cortex, from left to right, the amount of variance explained as each component is added is 26%, 48%, 60%, 67%, and 72%. In thalamus, the variance explained is 32%, 56%, 65%, 72%, and 77%.



(a) Cortical population CGF.

(b) Thalamic population CGF.

Figure 4.8: Population CGFs in the cortex and thalamus. (a) and (b): Cortex and thalamus, respectively. The main component of each figure is the population average CGF. Averaging over either the temporal (τ) or spectral (ϕ) dimensions, yields the line plots situated on the immediate right, or below the CGF. Error bars correspond to 2 standard errors.



(a) Average over all values of ϕ.

(b) Average over all values of τ.

(c) $\tau = 0$.

Figure 4.9: CGF statistics in cortex and thalamus. (a): The population CGFs have been averaged across all values of ϕ to yield these time-varying plots. Notice that the primary difference between the cortical and thalamic populations is evident here, in the extent of the delayed suppressive region within the CGF. (b): Here, an average has been taken across all values of τ to yield frequency-varying plots. Both the depth of the suppression, and the spectral extent, seem to be similar between brain areas. (c): The near-simultaneous region at $\tau = 0$. Another difference is somewhat evident here, in that the enhancement is restricted to only negative deviations in frequency within the thalamus, and both positive and negative deviations within cortex. This may simply be an averaging issue however (due to an under-representation in this particular dataset), since 4.6 (b) includes individual examples from the thalamic population where enhancement occurs at both deviations.

4.3.3.4 CGF Population Averages

In the context model, the dimensions of the CGF (given by τ and ϕ) are both *relative*. Thus, we can average the CGFs over an entire population of neurons in order to observe any contextual effects that are present on a grand scale (examples of such averaging were presented earlier, in figure 4.3). Figure 4.8 shows the average CGF structure in both cortical (a) and thalmic (b) populations. Averaging over either the temporal (τ) or frequency (ϕ) dimensions yields the line plots (provided with two standard error bars) located on the immediate right, and below, the averaged CGFs. The magnitude of these error bars further indicates that CGF structure is remarkably consistent throughout both populations of cells. For further comparison between the two areas, figure 4.9 provides the overlaid line plots for both the temporal and spectral averages (a) and (b), and the near-simultaneous region at $\tau = 0$ (c). Here, the effect of the near-simultaneity at $\tau = 0$ is not present within the temporal averages, purely because it is being averaged out with suppression at $\tau > 0$.

The delayed contextual suppression manifests itself as a large negative bump for values of $\tau < 0$, and seems to be present in both cortex and thalamus (albeit it on slightly different timescales). Such a region is likely to reflect aspects of the temporal analysis of sound; specifically, reported contextual effects such as forward suppression. Forward suppression has been detailed in the literature for decades (e.g. Calford and Semple (1995); Brosch and Schreiner (1997); Fitzpatrick et al. (1999); Bartlett and Wang (2005)) and is thought to relate to the psychophysically observed phenomenon of forward masking, an effect which is often ascribed to a cochlear mechanism (Moore, 1980; Jesteadt et al., 1982). One of the crucial differences however, is that forward suppression is typically associated with a reduction of excitation, in that a preceding sound will reduce the response to a future sound. The effect that we observe within the context model is more general than this. The delayed suppression here is more of a suppressive gain control. As a result of this, as well as reducing the effect of excitation (like forward suppression), it will also act to reduce the effect of inhibition.

One of the most notable differences between the CGF structure in cortex and thalamus is the temporal extent of this suppressive region. In the thalamus, the suppression reaches its minimum at 40-60 ms, before gradually lessening in impact (the suppressive effect is completely finished by 120-140 ms). Conversely, in cortex the suppressive region reaches its minimum at around 80-100 ms, and persists for a lot longer, not crossing zero until almost 200 ms.

The second of the contextual effects that was consistently observed is the presence of a near-simultaneous facilitatory band at $\tau = 0$. The functional significance of such a band is that contextual interactions between near-simultaneous tones seem to be different from non-simultaneous tones. Such a nonlinear simultaneous enhancement is something that has not been reported in the literature to any significant extent (but see Sadagopan and Wang (2009)). As a result of this, it may be a feature of complex sound processing (perhaps underlying harmonic analysis) that simply does not show under simple stimulus conditions, or simple analyses.

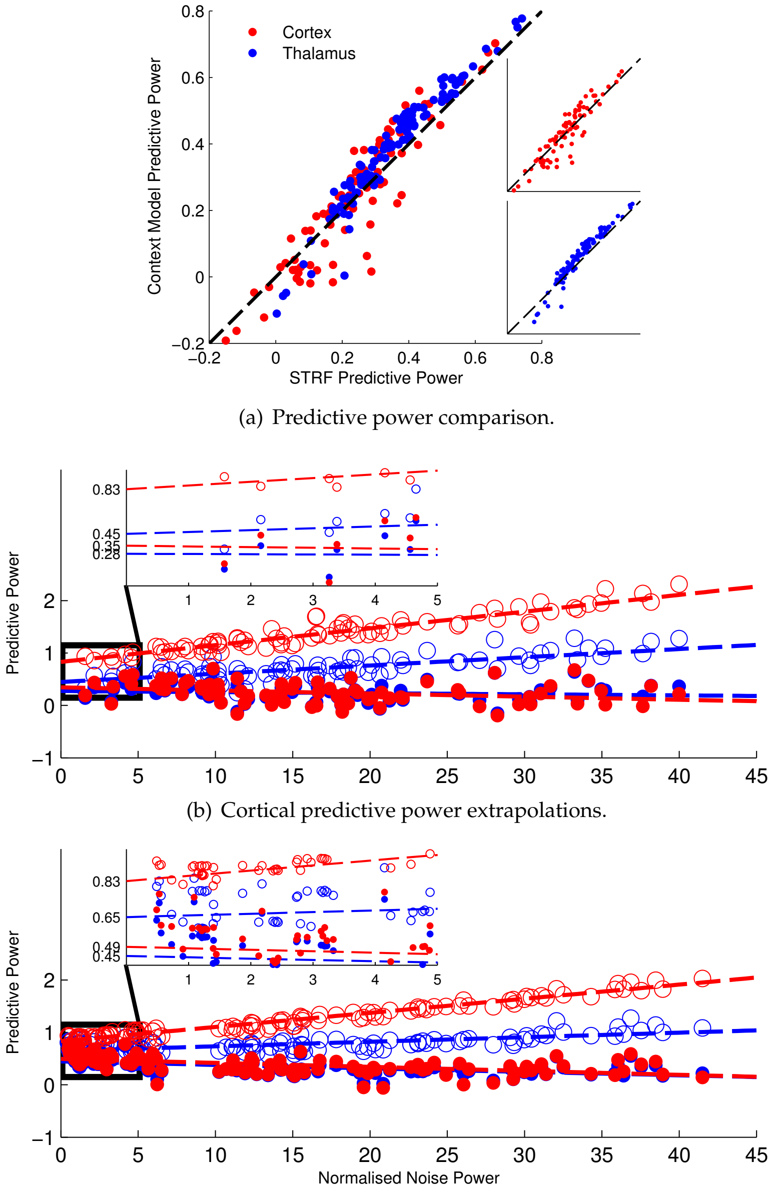
In the following section, we will evaluate the predictive capabilities of the single-CGF model, and demonstrate the significance of these structural regions.

4.3.4 Predictive Capability of the Single-CGF Context Model

The predictive power of this version of the context model is higher than reported previously for the fully-separated model with input nonlinearity (Ahrens et al., 2008a). In cortex, we extrapolated a lower bound predictive power of 0.35 (previously 0.32) and an upper bound predictive power of 0.83. This is in comparison to a linear model fit to the same data which yields an extrapolated lower bound of 0.27, and an upper bound of 0.44. The context model had greater predictive power in thalamus, with a lower bound of 0.50 and an upper bound of 0.83. In comparison, a linear model was able to achieve a lower bound of 0.44 and an upper bound of 0.64. These predictive power extrapolations are shown graphically in figure 4.10.

It is also worth noting that in calculating these lower bound values, a concious choice has to made in regards to what sort of function best describes the data to be extrapolated. In Ahrens et al. (2008a) a quadratic fit was utilised (yielding the aforementioned lower-bound predictive power of 0.32). Here, we have utilised a linear fit, yielding an extrapolation of 0.35, due to the fact that it is not clear in the current populations of data, whether a quadratic provides an accurate fit. Had a quadratic extrapolation been utilised for consistency, then the reported value of 0.35 actually increases to almost 0.4.

To further visualise the improvement that the context model provides, figure 4.10 (a) plots the cross-validated predictive power for both the context model and STRF model, in both cortex and thalamus. In both populations, the context model clearly provides the greatest predictive capability, indicated by the majority of points lying above the $y = x$ line.



(a) Predictive power comparison.

(b) Cortical predictive power extrapolations.

(c) Thalamic predictive power extrapolations.

Figure 4.10: Predictive capability of the single-CGF context model. (a): Cross-validated predictive powers for both the single-CGF context model and an STRF model are plotted against one another, for both cortex (red) and thalamus (blue). In thalamus, all but 8 cells lie above the $y = x$ line, indicating that the context model provides an improvement in the vast majority of thalamic recordings. There are slightly more cortical cells for which the context model does not provide a good fit. The distribution is skewed such that if an STRF model yields a poor predictive power, the context model typically performs worse. (b) and (c): Predictive power extrapolations for both cortex and thalamus. Context model fits are coloured blue, and STRF model fits are coloured red. Lower bound extrapolations are denoted by filled circles, and upper bound extrapolations are denoted by empty circles. The insets detail the point at which the extrapolation to zero noise power occurs, and the bounds themselves are provided.

There are several particularly interesting features here, that we wish to draw attention to.

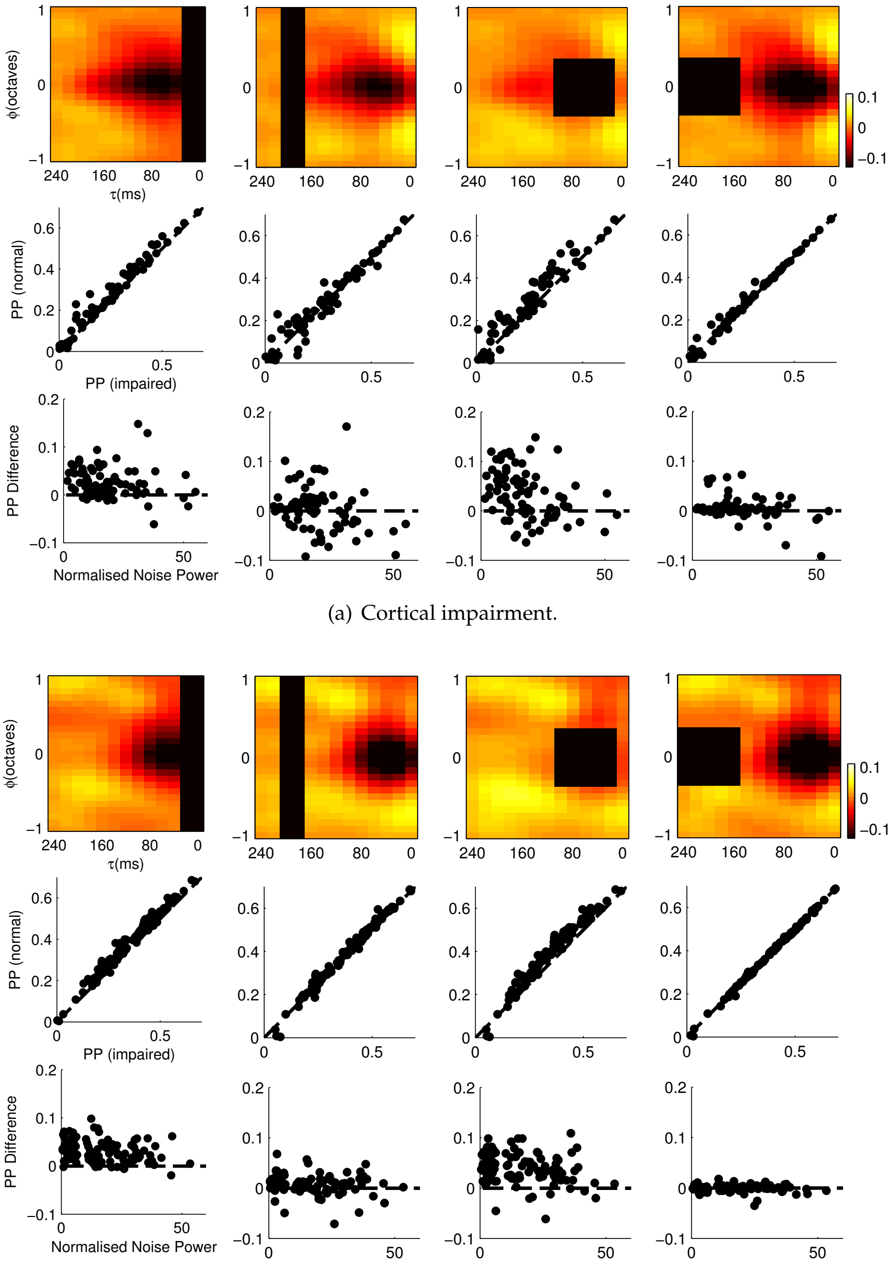
- Thalamus seems to be more linear than cortex. The extrapolated lower bound predictive power for an STRF model in thalamus is ~20% greater than that of cortex, indicating that a simple linear model is capable of capturing more stimulus-related variability in this sub-cortical structure.
- The gain from the context model is more significant in cortex than in thalamus. In cortex, we see a ~8% improvement over the linear model estimate whereas in thalamus this improvement shrinks to around ~4%.
- Almost all thalamic cells show an improvement through the use of the context model (for only 8 cells is it not the case). Interestingly, the cortical predictive power distribution takes on an interesting shape. It is somewhat skewed, such that if a simple linear model provides a poor fit to the cell, then the context model does even worse. If, on the other hand, a linear model provides a good fit to the cell, the context model always improves it. It seems that there are just some cortical cells that, even though they have a significant stimulus-related component in their neural response, are poorly fit by either a simple linear model or the context model. (It is worth noting, that a similar trend is also noticeable in the predictive power plots of Ahrens et al. (2008a), through the use of the fully separated context model).

We hypothesised that a primary reason for the context model to provide an increase in predictive power could be due to the fundamental nature of the structure present within the CGFs, and the contextual interactions that such structure represents. This is something that we specifically test in the next section.

4.3.5 Selective Impairment of CGF Structure

We were particularly interested in establishing what aspects of the context model were responsible for providing the improvement in predictive power. To that end, we chose to fit constrained (or "impaired") versions of the context model, in which different ranges of parameters within the CGF were not included within the estimation. We chose to focus on three specific cases. We wanted to

- Eliminate the simultaneous facilitatory structure at $\tau = 0$



(a) Cortical impairment.

(b) Thalamic impairment.

Figure 4.11: Selective impairment of CGF structure. (a) and (b): Cortex and thalamus, respectively. In both cases, the top row shows the population CGF, where the blacked out regions have been impaired (not included within the estimation). Columns 1 and 3 directly represent impairing the simultaneous and delayed structure, respectively, that is present in the vast majority of context model fits. Impairing this structure has a detrimental impact on predictive power (bottom two rows). Columns 2 and 4 represent controls, whereby the same number of parameters have been impaired as their structural counterparts. There is very little impact on predictive power in both control cases, indicating the importance of both simultaneous enhancement and delayed suppression in providing a predictive boost.

- Eliminate the delayed suppressive structure
- As two controls, eliminate regions devoid of "interesting" structure, where the regions were of the same size and shape as the previously defined impairments.

The results of this selective impairment of structure serve to highlight the fact that the structure we have presented within the population CGFs is what drives the predictive increase of the model. These results are shown in figure 4.11.

We started by impairing the first two columns of the CGF (corresponding to $\tau = 0$ and 1, at all relative frequencies ϕ (figure 4.11 (a)). The bottom two rows of (a) show the noticeable effect that this had on predictive power, causing a marked decrease in almost every cell, when the predictive powers of an unimpaired model are plotted against those of the impaired model. Additionally, the bottom row shows the difference between the unimpaired and impaired predictive powers, plotted as a function of noise power. Here, a positive number implies that impairing the given structure results in a loss of predictive ability.

Impairing the delayed region had a similar effect. The delayed region was defined to consist of a small window (running from $\tau = 2 \cdots 9$ and $\phi = -4 \cdots 4$, or 40-120 ms and 1/3 octaves around the current tone). Importantly, the number of parameters constrained in this impairment was almost identical to the previous simultaneous case. As with the previous impairment, the loss of this structure resulted in a drop in predictive power for almost every cell. This is true in both cortical and thalamic populations, and serves to indicate the potential importance of the structure that we observe.

However, in order to ensure that we were observing a genuine effect and to further validate that it was, in fact, the structure of interest that was responsible for the boost in predictive power, we carried out two controls. We chose to constrain areas of the CGF that were largely devoid of structure. In both cortex and thalamus, the delayed suppression tends to die out at ~160 ms, and thus we decided to "shift" the impairment zones into this region (shown in the second and fourth columns of figure 4.11. Both impairments had very little effect on predictive power. This is quite clearly indicated by the fact that the unimpaired predictive powers and impaired predictive powers are largely identical (falling onto the $y = x$) line in both cases).

These results serve to highlight the importance of precise combinations of spectrotemporal energy in proving the observable increase in predictive capability over a simple linear model.

4.3.6 Nonlinear Processing Characteristics of Cortical and Thalamic Subdivisions

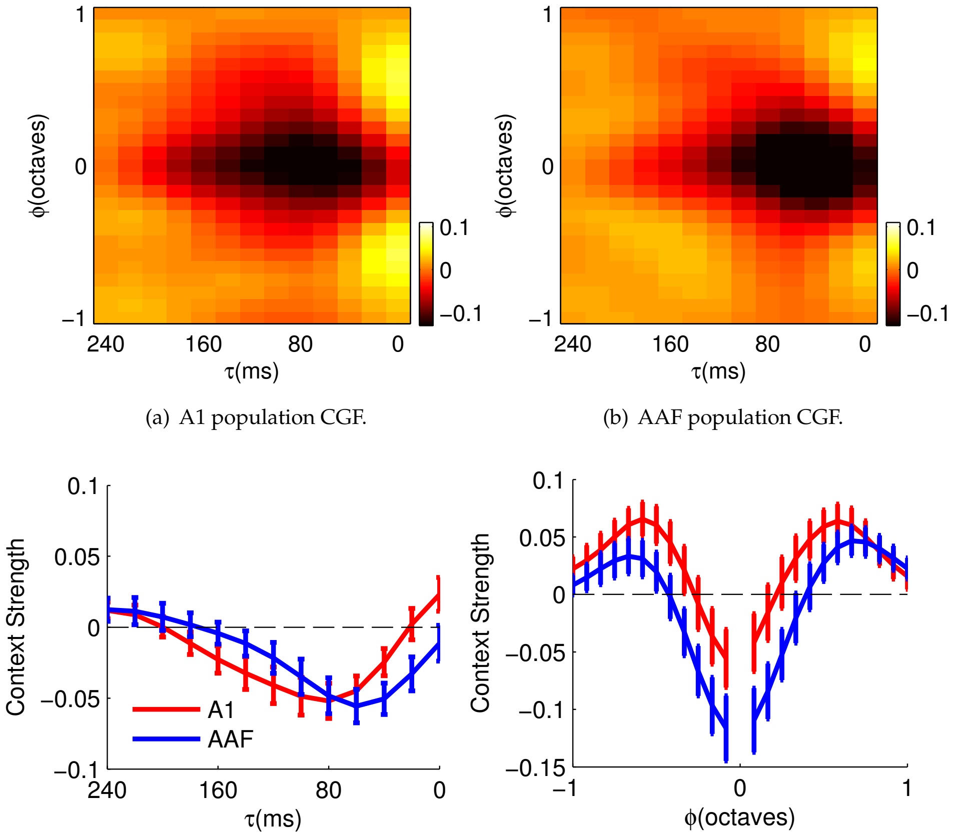
With the knowledge that the structure we observed within the CGFs was capable of providing a significant increase in predictive power, we wondered if these nonlinear contextual effects might differ between cortical areas or thalamic subdivisions.

4.3.6.1 A1 and AAF

Background material on cortical fields was provided in section 2.1.2.4.

Linden et al. (2003) described, in detail, differences in the STRFs fit to neurons in mouse A1 and AAF. One of the primary findings of this work is of a significant difference in temporal response properties between A1 and AAF. Peak latencies and receptive field durations of STRFs, and first spike latencies for responses to tone bursts were significantly longer in A1 than in AAF. In addition to this, there was significant overlap in the spectral properties of the two areas, but STRF bandwidths in A1 were very slightly broader than in AAF. Ultimately, these results serve to suggest that AAF may be specialised for faster temporal processing. Inspired by these findings, we wondered whether we could use the context model in an effort to identify differences in the way that these two auditory cortical fields process stimulus context, and how any differences relate to those discovered through the simpler STRF analyses. To this end, we analysed the CGFs that had been fit to neurons in the separate populations.

Figure 4.12 shows the population results of such an analysis. One of the most striking observations here is that the structure shared between the two cortical fields seems remarkably consistent. Both population CGFs include the delayed suppressive region, and the facilitatory strip at $\tau = 0$. Figure 4.12 (c) and (d) show the two key differences between the context model fits to these different cortical fields. In (d), we can observe that the magnitude of the simultaneous enhancement is slightly greater and also, the spectral extent of the interactions is very slightly broader in A1 than in AAF. In (c), we can clearly see that the timecourse of the temporal contextual interactions is faster in AAF than in A1. In AAF, the delayed suppression seems to reach a minimum at ~80 ms, whilst it takes A1 an additional 20-40 ms in order to reach the same suppressive depth. Once this minimum has been reached, AAF recovers quickly, and the effect of delayed suppression is largely over by ~180 ms. In A1 however, the extent of delayed suppression lasts until ~220 ms.



(a) A1 population CGF.

(b) AAF population CGF.

(c) Average temporal profile.

(d) Average near-simultaneous region ($\tau = 0$).

Figure 4.12: Nonlinear processing characteristics of areas A1 and AAF in the primary auditory cortex. (a) and (b): These figures show the CGF population averages in both A1 (left) and AAF (right). The structure looks remarkably consistent over both of these areas. (c): This shows the average temporal profile (averaged over ϕ). Notice that AAF exhibits a noticeably faster timecourse than A1. (d): This shows the spectral profile of the near-simultaneous region at $\tau = 0$. The profile in both areas are similar, although A1 seems slightly more spectrally broad.

We also analysed the PRFs from the context models, and compared them to STRFs in order to assess the effects of including the CGF component within the model. Following Linden et al. (2003), we extracted a number of different statistics from both the STRF and PRF population (summarised in figure 4.13. The *peak latency* was the time to the center of the peak in the first subfield of the receptive field (this was usually an excitatory subfield, but occasionally an inhibitory subfield). *Receptive field duration* was defined to be the time from the beginning of the first subfield to the end of the last subfield. *Excitatory subfield width* was defined to be the width at half-maximum of the positive peak in the temporal profile of the receptive field, whilst the *inhibitory subfield width* was defined to be the width at half-minimum of the negative peak in the temporal profile.

	A1	AAF
STRF Peak Latency (ms)	43.85 ± 1.03	24.88 ± 0.50
STRF Receptive Field Duration (ms)	159.49 ± 2.94	125.12 ± 2.10
STRF Excitatory Subfield Width (ms)	51.28 ± 1.20	36.74 ± 1.35
STRF Inhibitory Subfield Width (ms)	94.87 ± 1.64	82.79 ± 0.92
STRF Excitatory Bandwidth (kHz)	20.92 ± 0.47	17.18 ± 0.74
STRF Inhibitory Bandwidth (kHz)	16.90 ± 0.36	15.25 ± 0.46
STRF Normalised Excitatory Bandwidth	1.19 ± 0.03	0.92 ± 0.02
STRF Normalised Inhibitory Bandwidth	1.02 ± 0.02	0.91 ± 0.02
PRF Peak Latency (ms)	40.77 ± 0.77	29.07 ± 0.91
PRF Receptive Field Duration (ms)	163.08 ± 3.28	146.98 ± 2.28
PRF Excitatory Subfield Width (ms)	50.77 ± 1.07	39.53 ± 1.24
PRF Inhibitory Subfield Width (ms)	105.64 ± 2.24	100.47 ± 1.43
PRF Excitatory Bandwidth (kHz)	19.75 ± 0.40	16.38 ± 0.65
PRF Inhibitory Bandwidth (kHz)	15.38 ± 0.45	15.11 ± 0.47
PRF Normalised Excitatory Bandwidth	1.18 ± 0.03	0.91 ± 0.02
PRF Normalised Inhibitory Bandwidth	0.95 ± 0.03	0.91 ± 0.02

Table 4.1: Spectral/temporal profile differences between A1 and AAF.

The *excitatory* and *inhibitory bandwidth* was defined to be the width at half-maximum (or half-minimum) of the positive (or negative) peak in the spectral profile. The *normalised bandwidth* was achieved by normalising either excitatory or inhibitory bandwidths by the *best frequency* of the receptive field (defined to be the frequency corresponding to the maximum in the spectral profile). Table 4.1 summarises these statistics for both the STRF and PRF populations within A1 and AAF. Moreover, figure 4.14 summarises these statistics graphically over the entire cortical and thalamic populations.

The first thing to immediately note is that the STRF statistics confirm what was reported by Linden et al. (2003). Namely, that the inhibitory subfield width is shorter in AAF that in A1, and the spectral bandwidth is slightly greater in A1. Interestingly, there is one particular statistic in which there exists a significant difference between the PRF and STRF populations. The inhibitory subfield widths in the PRFs are slightly longer than observed in the STRFs. As a result of this, the overall receptive field durations are typically a little longer. This is best observed in figure 4.14, where inhibitory subfield width, and receptive field duration, are the only statistics can be seen to deviate away from $y = x$.

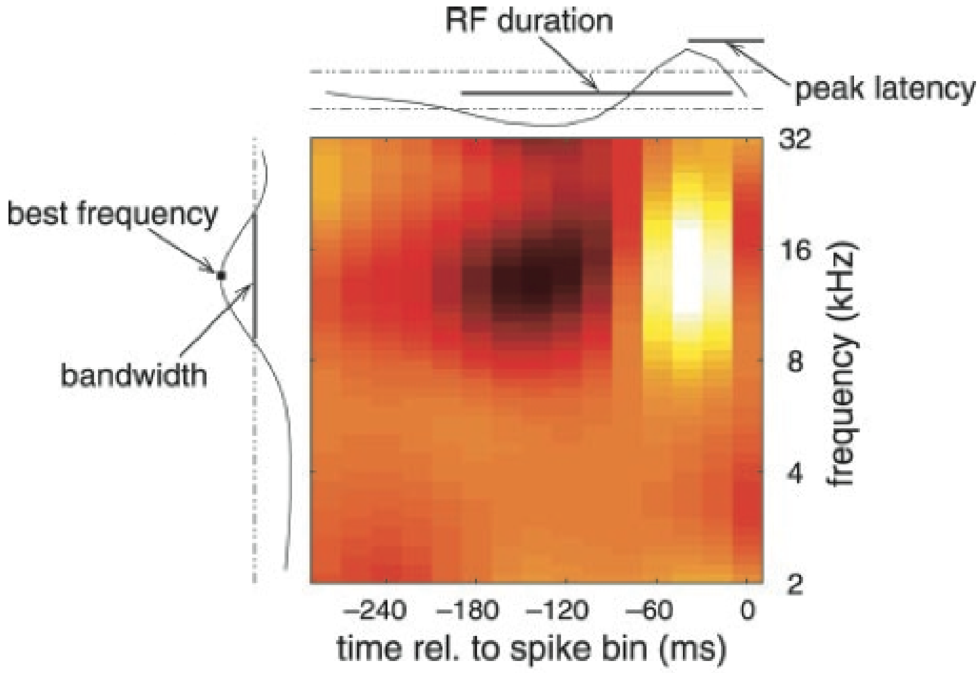
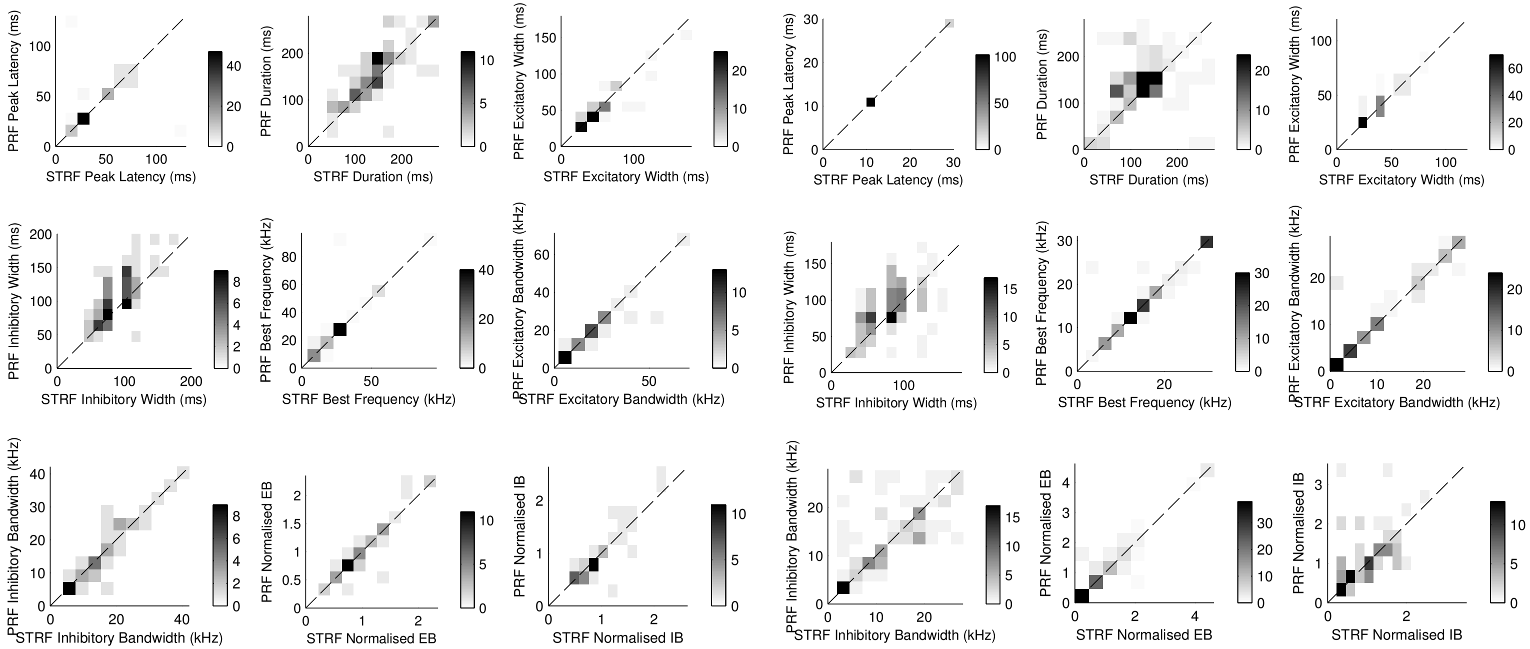


Figure 4.13: Receptive field statistics (adapted from Linden et al. (2003)). The *peak latency* was the time to the center of the peak in the first subfield of the receptive field (this was usually an excitatory subfield, but occasionally an inhibitory subfield). *Receptive field duration* was defined to be the time from the beginning of the first subfield to the end of the last subfield. *Excitatory subfield width* was defined to be the width at half-maximum of the positive peak in the temporal profile of the receptive field, whilst the *inhibitory subfield width* was defined to be the width at half-minimum of the negative peak in the temporal profile. The *excitatory* and *inhibitory bandwidth* was defined to be the width at half-maximum (or half-minimum) of the positive (or negative) peak in the spectral profile. The *normalised bandwidth* was achieved by normalising either excitatory or inhibitory bandwidths by the *best frequency* of the receptive field (defined to be the frequency corresponding to the maximum in the spectral profile).



(a) Cortex.

(b) Thalamus.

Figure 4.14: Comparing PRFs and STRFs: Spectral and temporal profile measures. Here, a selection of spectrotemporal differences between the STRF and PRF populations are shown. The colour map explicitly refers to density, and shows the number of cells within the given bin. For the vast majority of measures, STRFs and PRFs are remarkably similar. The main variation occurs in inhibitory subfield width, which also causes a difference in receptive field duration. Here, PRFs are noticeably longer. This is clearly observable in the off-diagonal terms in the relevant plots.

Linear models are capable of capturing suppressive stimulus effects through negative regions within their spectrotemporal profiles. The addition of the CGF to a linear framework provides an additional way to create such effects. Ahrens et al. (2008a) showed that, in comparing the amount of relative suppression between two multilinear models, that a model devoid of context had significantly more relative suppression in its linear component that a model with contextual information. This suggests that the contextual component of the context model can be used to explain some of the suppression present within simpler linear estimates (attributing such suppression to the nonlinear effects of stimulus context). We replicated this analysis, and directly compared the amount of relative suppression (given by min(filter)/max(filter) − min(filter)), in both the STRF and PRF populations. The results are presented in figure 4.15.

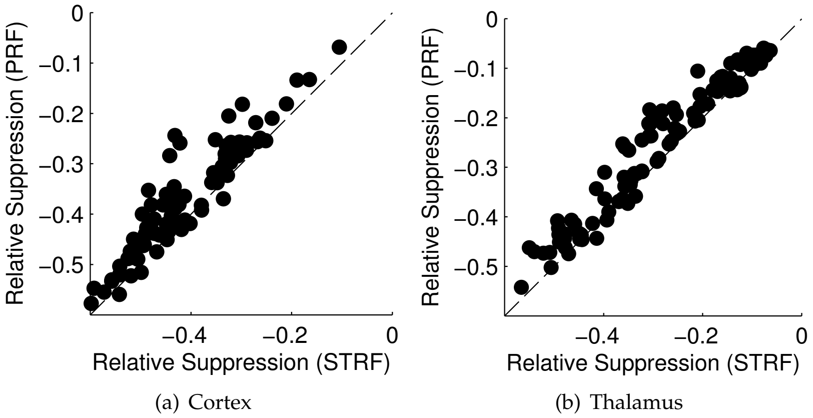
Figure 4.15 clearly shows that the amount of relative suppression is greater within the STRF population, for both cortex and thalamus. This in an important insight, and provides us with interpretation as to why the temporal structure of the CGFs is different between A1 and AAF.

Linden et al. (2003) showed that the inhibitory subfield width in A1 was significantly different from AAF. This is also shown here, in table 4.1. An interesting observation however is that this temporal difference does not seem to be as obvious within the PRF population (a difference of ~5 ms, as opposed to ~14 ms in the STRFs). This is actually quite an important distinction, since it shows that the temporal difference originally observed within the STRF population is better captured by the contextual component of the mutilinear framework. This indicates that the difference in temporal processing between A1 and AAF can be attributed (in part, at least) to the effects of stimulus context.

Ultimately, these results show that differences in the temporal processing abilities of A1 and AAF may reflect, in part, multiplicative stimulus interactions, such as forward suppression.

4.3.6.2 Subdivisions of the Medial Geniculate Body

Given that we were able to use the context model to shed light on nonlinear differences between cortical fields, we were curious as to whether we could also identify any similar differences within the auditory thalamus. Although the properties of lemniscal thalamic receptive fields have been studied (Miller et al., 2002), the responses of the MGB subdivisions to complex sounds, and the effect of (short-term) acoustic context remains elusive (although see Wehr and Zador (2005), for thalamic intracellular forward



(a) Cortex

(b) Thalamus

Figure 4.15: Comparing PRFs and STRFs: relative suppression. We directly compared the amount of relative suppression (given by min(filter)/max(filter) − min(filter)) in both cortex and thalamus. The trend in results was similar in both areas, whereby in almost all cells, the amount of relative suppression is greater within the STRF than the corresponding PRF. This suggests that the CGF component of the context model is better able to account for the suppressive effects of stimulus context.

suppression). Given this, we chose to subdivide our thalamic population (using histological delineation (see figure 4.16)), and analysed the context model fits to the resultant cells.

Background material on the subdivisions of the thalamus was provided in section 2.1.2.3.

We first analysed the differences between the STRF and PRF populations, which were shown earlier in figure 4.14. The same trends present within the cortical recordings are also present within the thalamus, in that the primary difference between populations is in the inhibitory subfield width and, as a result, receptive field duration (differences between the populations can be seen in any off-diagonal elements within figure 4.14). The difference in relative suppression that was discussed earlier, also holds (figure 4.15).

Table 4.2 shows the thalamic statistics, broken down by the different subdivisions. The first thing to note is that the main way in which these statistics vary from cortex is temporally. Peak latency, receptive field duration, and subfield widths are all shorter in thalamus than in cortex (these temporal differences also conform to trends observed by Miller et al. (2002)).

Figure 4.17 (top row) shows the population averaged CGFs, for the different subdivisions. They all share similar structure, indicating that there does not seem to be a significant difference in the nonlinear processing characteristics of the different subdi-

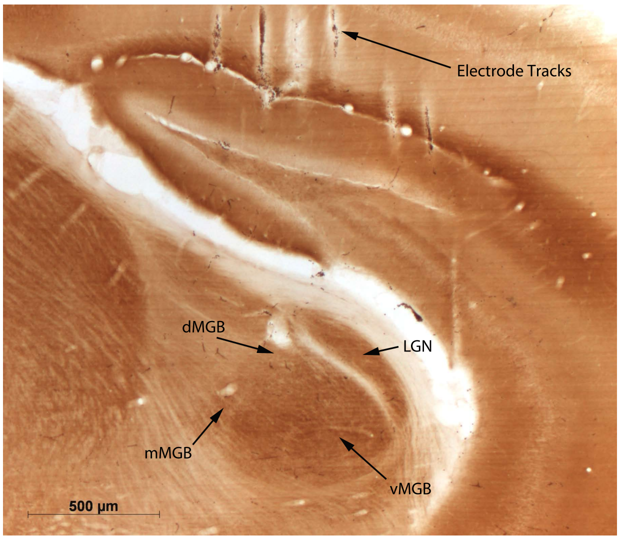


Figure 4.16: An example of thalamic histology. This representative 50 μm slice shows a cytochrome oxidase stained MGB, with the relevant subdivisions clearly marked. In addition to the auditory part of the thalamus, the lateral geniculate nucleus (LGN), part of the visual thalamus, can also be observed. At the top of the slice, damage left by the electrode array as it was pushed through cortex can be observed.

	vMGB	mMGB	dMGB
STRF Peak Latency (ms)	13.14 ± 0.29	14.71 ± 0.51	15.45 ± 1.70
STRF Receptive Field Duration (ms)	129.41 ± 2.37	114.12 ± 3.19	110.91 ± 4.98
STRF Excitatory Subfield Width (ms)	34.51 ± 0.74	35.29 ± 0.96	23.64 ± 2.19
STRF Inhibitory Subfield Width (ms)	81.57 ± 1.12	73.53 ± 1.58	69.09 ± 2.98
STRF Excitatory Bandwidth (kHz)	12.84 ± 0.42	11.27 ± 0.57	5.96 ± 0.78
STRF Inhibitory Bandwidth (kHz)	13.85 ± 0.33	12.31 ± 0.45	11.03 ± 0.36
STRF Normalised Excitatory Bandwidth	1.17 ± 0.05	0.78 ± 0.04	0.41 ± 0.05
STRF Normalised Inhibitory Bandwidth	1.09 ± 0.02	0.89 ± 0.03	0.83 ± 0.05
PRF Peak Latency (ms)	13.14 ± 0.29	14.71 ± 0.51	15.45 ± 1.70
PRF Receptive Field Duration (ms)	140.78 ± 2.37	125.29 ± 3.70	156.36 ± 11.13
PRF Excitatory Subfield Width (ms)	33.73 ± 0.53	37.65 ± 1.29	23.64 ± 2.19
PRF Inhibitory Subfield Width (ms)	84.71 ± 1.43	76.47 ± 1.89	89.09 ± 3.40
PRF Excitatory Bandwidth (kHz)	13.35 ± 0.41	12.41 ± 0.56	5.96 ± 0.78
PRF Inhibitory Bandwidth (kHz)	14.72 ± 0.32	13.31 ± 0.42	15.37 ± 1.04
PRF Normalised Excitatory Bandwidth	1.22 ± 0.04	0.83 ± 0.04	0.41 ± 0.05
PRF Normalised Inhibitory Bandwidth	1.23 ± 0.03	0.93 ± 0.03	1.18 ± 0.10

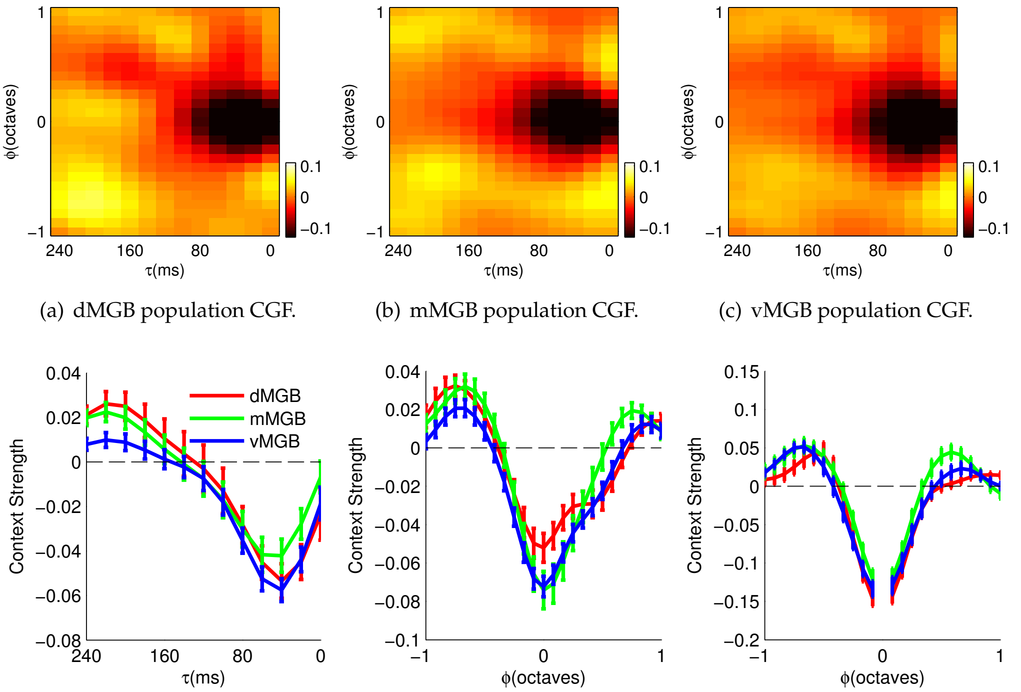
Table 4.2: Spectral/temporal profile differences between thalamic subdivsions.

visions. The bottom row of the figure shows temporal and spectral averages, and the near-simultaneous region, from each of the population CGFs. Again, the overlap is significant.

An important point to remember is that even though there is great overlap between these different divisions, the context model still provides an increase in predictive power, for almost every cell in the population, over an STRF model. This increase does not take place if the structure within the CGF is impaired in some way (see section 4.3.5). Thus, it seems likely that the delayed suppression and near-simultaneous enhancement is still providing insight into biological mechanisms present within all three subdivisions.

One important point here is to do with data set size. The number of cells that were able to be accurately assigned to the dorsal subdivision is small, consisting of only 11 (as opposed to 34 in mMGB, and 51 in vMGB). This is likely due to recording bias, in that vMGB is the largest (and thus, easiest) subdivision to target experimentally. Although dMGB is likely to be targeted in all penetrations (since it lies directly above vMGB), it proved difficult to elicit responses from. It has, however, been previously reported that dMGB does not respond particularly well to simple acoustic stimuli, such as clicks, which we used as search stimuli in our experiments (Buchwald et al., 1988). This is also a contributing factor to the size of the dMGB error bars that are provided in table 4.2.

The fact that no differences were observed between the ventral and medial subdivisions is puzzling however, especially given the fact that mMGB has been implicated in long-term acoustic adaptation, through a variety of stimulus-specific adaptation studies (Anderson et al., 2009b; Antunes et al., 2010). One potential explanation is simply that of definition. Here, stimulus context is explicitly defined to be a short-term window that surrounds each tone within the stimulus spectrogram. This is different from the more long-term context that is typically associated with the aforementioned change detection studies. With this in mind, it may genuinely be the case that biological mechanisms, such as forward suppression, are in fact similar across thalamic subdivisions, and the context model is correct in its invariant structure. This however, is a hypothesis that remains to be tested.



(a) dMGB population CGF. (b) mMGB population CGF. (c) vMGB population CGF.

(d) Average temporal profile. (e) Average spectral profile. (f) Near-simultaneous region ($\tau = 0$).

Figure 4.17: Comparison between the three thalamic subdivisions. (a), (b), and (c): These plots show the population CGFs for the different thalamic subdivisions. All three CGFs are incredibly similar in structure. (d), (e), and (f): These plots show the averaged temporal and spectral profiles, and the near-simultaneous region, present within the above CGFs. The level of overlap between the different subdivisions is clear.

4.4 Discussion

4.4.1 Nonlinear Modeling of Neural Responses

We have presented a variant of the context model (Ahrens et al., 2008a), that does not utilise an input nonlinearity, and includes two inseparable receptive fields. One of the primary assumptions that underlies the context model is the notion of contextual invariance; that is, the contextual effects in the model are identical for all frequencies and time-lags. Before proceeding to analyse such a model, we wished to establish whether this was a valid assumption to make. In order to do this, we described a framework for "splitting" the context model, where multiple CGFs can be associated with different underlying elements within the PRF. For the vast majority of splits that we tried, structure across CGFs was typically similar, and predictive power decreased when compared directly to the single CGF model. The only split in which this was not entirely obvious, was when contextual effects were split across the excitatory and inhibitory components of the PRF. In this case however, the similarities clearly outweighed the dissimilarities, especially for cells with a high predictive capability. Ultimately, our results indicated that contextual invariance was a valid assumption to make, and we proceeded with analysis of the single CGF model accordingly.

One of the most striking results of this study is that the single CGF context model consistently exhibits structure within the CGF weights, indicative of a near-simultaneous enhancement (facilitatory interactions at $\tau = 0$), and a delayed dampening effect. As a result of this structure, the model outperforms standard linear estimates of neural firing in both cortex and thalamus. Interestingly, through this analysis we were also able to show that the thalamus (as one might expect) is somewhat more linear than cortex. Through the use of a simple linear model, we were able to capture between 40% and 60% of the stimulus-related variability.

As a final point, we were interested to see whether we could use the context model to identify differences in nonlinear processing characteristics within different cortical and thalamic subdivisions. In the cortical data, this was certainly the case; the delayed suppression present within the CGFs when the model was fit to data from either A1 or AAF differed in time course. There was no significant difference in inhibitory subfield width within the linear component of the context model, whereas there was a significant difference within STRF fits to the same data. Ultimately, this suggests that the context model has been successfully able to attribute inhibitory changes within a simple linear model to the nonlinear effects of stimulus context. There were no significant differences between context model fits to multiple subdivisions of the thalamus. This result actually provides testable predictions for future experiments. To our knowledge, even classical two-tone paradigms have not been carried out in the thalamus (although see the intracellular work of Wehr and Zador (2005)), and what results have been established have not been attributed to a particular subdivision. In short, the mechanisms underlying contextual interactions within the multiple subdivisions of the auditory thalamus remain an open question.

It is certainly worth noting however, that the approach taken within this thesis is primarily "black-box". That is, the modelling framework only acts upon an input (a spectrogram) and an output (a spike train). Everything in between is treated as a unknown. Such a level of abstraction provides limitations on how best to interpret results. Here, contextual effects are interpreted as nonlinear stimulus interactions. Where these contextual effects arise however, be it at the level of the periphery, or further downstream, is a question that still needs to be addressed.

4.4.2 Mechanisms of Stimulus Context

The inclusion of an inseparable local context has not only allowed us to predict both cortical and thalamic responses more accurately than previously reported, but has allowed us an insight into the potential mechanisms underlying the nonlinear effect of stimulus context. Stimulus context is something that has been studied for decades in the literature. Early work was primarily focussed on studying extracellular responses to two successive simple sound stimuli. Typically, such an experiment would study the influence of a preceding stimulus on a subsequent stimulus; essentially, a neural equivalent of the tone-masker paradigms used frequently in psychoacoustics. Stimulus context is capable of eliciting two primary effects on the neural response. The first effect is a facilitatory interaction, whereby the second stimulus in a pair elicits enhanced firing. Such facilitatory responses have been found in a variety of species including anesthetised cats (Brosch and Schreiner, 2000), birds (Margoliash and Fortune, 1992), macaque (Brosch et al., 1999), and awake marmoset (Bartlett and Wang, 2005). The second effect is a suppressive interaction, whereby the second stimulus in a pair acts to actively suppress the neural response. Such suppressive effects have also been observed in a variety of species including anesthetised cats (Calford and Semple, 1995; Brosch and Schreiner, 1997), awake rabbits (Fitzpatrick et al., 1999), and marmosets (Bartlett and Wang, 2005). More recently, several studies have directly address the synaptic and cellular mechanisms that give rise to such contextual effects (Wehr and Zador, 2003, 2005; Tan et al., 2004). The nature of the contextual effects that we present are largely consistent with these studies. Moreover, the context model is explicitly able to show how such contextual effects impact upon neural responses to complex sounds.

4.4.3 Implications for Past and Future Linear Analyses

A standard linear model is not capable of capturing a large amount of the stimulus-related variability in either cortex or thalamus, suggesting that a different approach is needed. The addition of contextual terms to such a model provides an increase in predictive capability, but also allows for a novel interpretation of previous STRF analyses. We showed that the linear component of the context model typically contains less relative suppression than that of the corresponding STRF. Since the CGF provides an additional way to model suppressive effects, this suggests that some of the suppressive regions observed through traditional STRF analyses can be better explained through the use of the context model. This is an important point because it provides a means of establishing which suppressive effects observed in the traditional linear analysis are actually directly attributable to the multiplicative effects of stimulus context. We showed that this seems to be the case for data recorded in A1/AAF, and that the difference in STRF inhibitory timecourse reported by Linden et al. (2003) seems to be due to differences in the timescale of delayed nonlinear suppression. As a result of this, we would hope that such a multilinear analysis could be successfully applied to a multitude of other STRF studies, in an attempt to explicitly tease out differences that can be attributed to nonlinear contextual interactions.

Ultimately, we have provided a model framework whose ability lies in the successful estimation of nonlinear interactions from the neural responses to complex sounds. Such a framework provides a novel extension to the study of receptive fields in multiple brain areas, and extends existing understanding of the way in which stimulus context drives complex auditory responses.

V

Nonlinear Sensitivities to Stimulus Context in Diverse Acoustic Environments of Increasing Complexity

Outline

This chapter is the second of two primary results chapters within this thesis. Previously, we showed that we could use the multilinear framework to successfully estimate nonlinear interactions from the neural responses to complex sounds. Here, we use a stimulus that varies in spectrotemporal density, in order to mimic a range of diverse acoustic conditions. We analyse the neural responses to such complex stimuli, and use the context model to investigate the underlying linear and nonlinear mechanisms.

5.1 Introduction

The mammalian auditory system has a remarkable capacity to operate in widely diverse acoustic environments, from the relative silence of a quiet library to the dense acoustic ambience of a Pearl Jam concert. How the auditory system is capable of achieving such a daunting task is not well understood. We wish to investigate the nonlinear mechanisms that contribute to such behaviour.

In order to address this question, we recorded extracellularly from the auditory cortex and thalamus of anaesthetised mice during presentations of spectrotemporally-rich dynamic random chord (DRC) stimuli. In contrast to the version of the DRC stimulus used in experiments described in chapter 4, here we used a DRC with a switching structure, whereby the spectrotemporal density (defined in terms of number of tone pulses per octave) regularly changes, thus mimicking a range of diverse acoustic conditions. We first present an analysis of the neural responses themselves, and quantify how the variability within the responses changes as a function of the stimulus environment. We then proceed to analyse the effects of contextual dependence; that is, we specifically examine whether or not the neural response to a specific stimulus can be influenced by the stimulus preceding it.

Finally, we show how we can use both linear spectrotemporal receptive field (STRF) models, and the extended context model framework (discussed at length in the previous chapter) to both predict responses to complex sounds, and elucidate the role of nonlinear contextual interactions in diverse acoustic environments.

5.2 Materials and Methods

5.2.1 Animals

Four adult CBA/Ca mice were used to collect the cortical data and five adult CBA/Ca mice (6-8 weeks old) were used to collect the thalamic data. In addition, five CBA/Ca mice from an earlier control study were also used to collect thalamic data under a different anaesthetic.

5.2.2 Experimental Procedures

Surgical, histological, and electrophysiological procedures were as described in the previous chapter. The previous control data were recorded from mice that had been anaesthetised with a cocktail of urethane and bupranorphine, rather than the ketamine and medetomidine protocol described previously. All other procedures were identical.

5.2.3 Stimulus

In the previous chapter, we utilised a spectrotemporally-rich complex sound, known as a dynamic random chord (DRC). Such a stimulus consists of multiple cosine-gated 20ms tone pulses that make up a random chord at each point in time. Here, we use a DRC with a variety of modifications. A single trial of this modified DRC lasts for 90 seconds (30 seconds longer than the previous version of the stimulus). Importantly, this modified DRC consists of three different spectrotemporal densities (given constraints on the amount of data required to adequately estimate statistical models, more than three densities would have proved difficult); sparse (0.5 tone pulses/bin/octave), mid (1 tone pulses/bin/octave), and dense (2 tone pulses/bin/octave). The centre frequencies of the tone pulses were drawn from 24 different possibilities, ranging from 8-32 kHz. Thus, the spectrotemporal densities can also be given per chord, with the sparse density yielding (on average) 1 tone pulse/bin/chord, mid yielding 2 tone pulses/bin/chord, and dense yielding 4 tone pulses/bin/chord. Each tone pulse is played at 55 dB SPL.

A 20 ms tone pulse length was used, with a 5 ms clock. This has the effect of "jittering" the tones, such that the onset of a new tone can occur whilst other tones are still being presented. In order to prevent overlap within frequency bands, each tone was given a pseudo-refractory period, such that at a given frequency, multiple tones cannot be played simultaneously.

A specific switching structure was also imposed onto the stimulus. Over the course of the 90 second trial, the density switches every 3 seconds. The switching order was designed in such a way that every density is preceded by every other density (including its own) at least once. The stimulus was repeated for 20 trials leading to a presentation time of 30 minutes. Examples of these different densities, and the switching structure, are provided in figure 5.1.

5.2.4 Modelling Neural Responses to Sound

Much of the modelling that is utilised here was discussed at length in chapter 3 and utilised in chapter 4.

Briefly, we fit both linear and multilinear models to the DRC-evoked neural responses. The STRF model was discussed in section 2.2.2, and the stimulus-response function (the function relating the stimulus spectrogram to the neural response) was given by equation (2.1). Estimation of the STRFs was carried out using the automatic smoothness determination algorithm (ASD) algorithm, due to Sahani and Linden (2003a). This technique was discussed in section 3.4.4.

The mathematical details of the multilinear framework that we utilise here was the focus of chapter 3. We use the model described by equation 3.24. Estimation was carried out using the alternating least squares (ALS) procedure, discussed in section 3.4.1.

5.2.5 Neuronal Populations

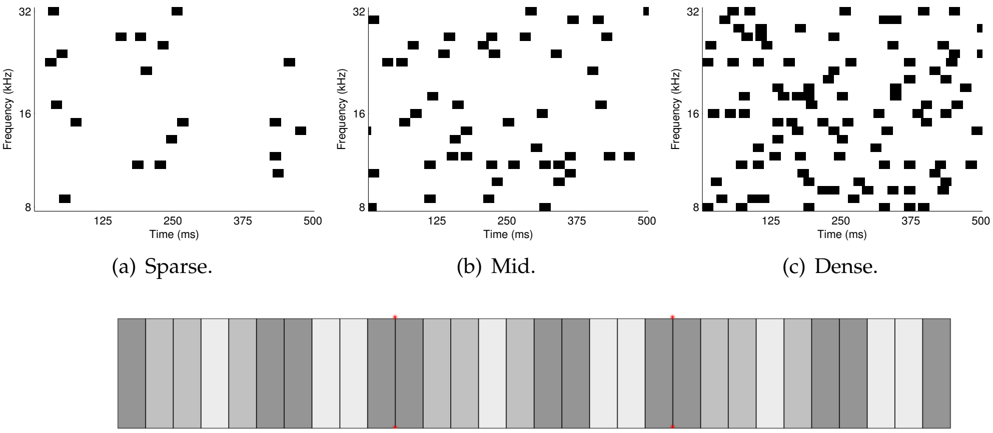
We used the signal power metric of (Sahani and Linden, 2003b) to establish which of our neuronal recordings exhibited a significant amount of stimulus-related variability, and were worth further analysis. We discarded all recordings that did not have a signal power at least 1 standard deviation away from zero. This left us with populations of neuronal responses to the switching dynamic random chord stimuli recorded in 46 cortical sites and 83 thalamic sites.

The data are pooled over multiple cortical fields, and different thalamic subdivisions. For this particular study, we were interested in evaluating differences over multiple stimulus conditions, and splitting the data up into subgroups would result in a loss of statistical power within each group.

5.3 Results

5.3.1 Reliability of Neural Responses to Different Densities

The DRC stimulus is specifically tailored towards estimating statistical models of neural responses. However, we were particularly curious to establish what changes (if any) were present within the neural responses themselves as the spectrotemporal density



(a) Sparse.

(b) Mid.

(c) Dense.

(d) Switching structure.

Figure 5.1: Switching DRC stimulus. (a), (b), (c): 500 ms samples of the different DRC densities used to construct the full stimulus. The density of the sparse stimulus is (on average) 0.5 pulses/bin/octave, the mid stimulus is 1 pulse/bin/octave, and the dense stimulus is 2 pulses/bin/octave. Our bin width is 5 ms. (d): The switching structure of the stimulus. Here, colour is used to denote density, with the lightest shade representing sparse, and the darkest shade representing dense. Every stimulus block lasts for 3 seconds, and the full duration of the stimulus is 90 seconds. The switching structure is specifically designed such that every density is preceded by every other density (including its own) at least once. The stimulus is also designed to contain repetition in stimulus blocks, such that multiple presentations of an identical stimulus can be gathered. The red dots placed at at stimulus blocks 10 and 20, denote a change in the "token" used. This means that, until block 10, all sparse stimulus blocks are identical (as are the other densities). This repeats every 10 blocks such that there are three different presentations of each density.

of the stimulus changed. Interestingly, simply by looking at the data themselves, one can notice specific changes in the way that the individual neurons in the populations respond. It seems to be quite clear that the most notable difference is in regards to the inherent variability across multiple presentations of the same stimulus. A specific example is presented in figure 5.2. Here, for a representative thalamic neuron, we show the responses to three, 3 second segments of the different densities included within the stimulus. The colour code is set up such that as the stimulus density increases, the colour darkens.

In figure 5.2 *top row*, we can see the raster plots for twenty presentations of the sparse DRC. Clearly, this stimulus seems to evoke a particularly reliable response, indicated by the vertical striping within the rasters. That is; this neuron responds in a similar fashion to the same features within the stimulus on each trial. It seems to have quite low trial-to-trial variability. As we move to the mid stimulus on the row below however, we can observe that the variability seems to have increased slightly in response to a slightly denser stimulus. There is still some vertical striping, indicating that neuron is still responding at least somewhat reliably to the stimulus, but it is certainly not as pronounced as before. The same trend is present as we move onto the third row. In response to the densest of the three stimuli, the trial-to-trial variability of the neural response seems to have increased dramatically, and there is no longer any clear indication that the neuron is responding reliably to particular features within the stimulus.

This trend, that trial-to-trial variability increases as a function of density, is something that seems to persist over the entire population of both cortical and thalamic responses to the same stimulus. In order to try and quantify this change in neural reliability, we chose to utilise a statistical approach, and calculate explicitly, the amount of stimulus-related variability within the data. To do this, we used the signal power estimator, proposed by Sahani and Linden (2003b). This technique was designed around the principle that neural responses are noisy, and difficulties can arise from the fact that repeated presentations of the same stimulus can elicit variable responses. The estimator itself is based around the segregation of response power (where we use the term power here, to denote variance over time) into a stimulus-dependent *signal* component and a *noise* component. This stimulus-dependent component, the signal power, can be estimated as

$$\hat{P}(\mu) = \frac{1}{N-1}(NP(\overline{\mathbf{r}^{(n)}}) - \overline{P(\mathbf{r}^{(n)})}) \tag{5.1}$$

where $P(\overline{\mathbf{r}^{(n)}})$ and $\overline{P(\mathbf{r}^{(n)})}$ are both trial averaged quantities denoting the power of the average response, and the average power per response, respectively. Subtracting this expression from $\overline{P(\mathbf{r}^{(n)})}$ yields an expression for the averaged noise power

$$\hat{P}(\eta) = \overline{P(\mathbf{r}^{(n)})} - \hat{P}(\mu) \tag{5.2}$$

This signal power measure enables us to both quantify the observed change in reliability and ask if the observation made from individual neuron examples is true across the entire population.

Reliability over trials can be directly related to stimulus-related variability over time, in the sense that the more repeatable a neural response is (the more vertical striping is present within the raster plots), the more variability will be present over time. This is simply due to the large deviation from the mean firing rate, caused by the reliable spiking activity. Conversely, if the reliability across trials is relatively low, then this will result in low variability over time, due to the lack of large deviations from the

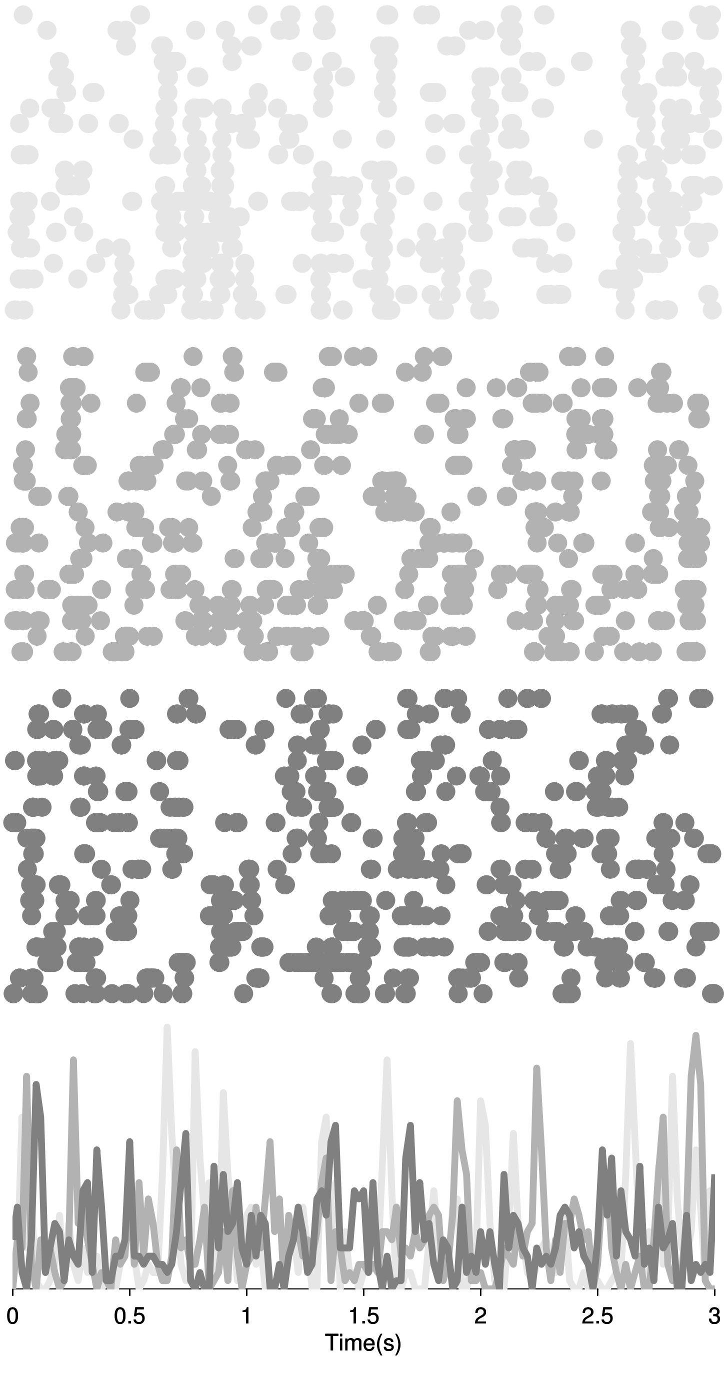
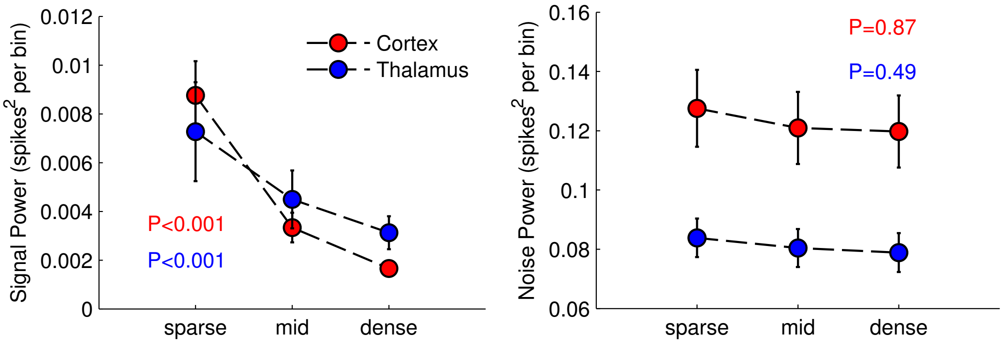


Figure 5.2: Reliability of neural responses to different densities. Here, we show 3 second responses (one complete stimulus block) from a representative thalamic recording. The colours of the raster plots are as defined previously, and correspond to the different densities (light - dark, sparse - dense). The PSTHs of each raster are overlaid at the bottom. The primary result to note, is that trial-to-trial variability increases as a function of density.



(a) Signal power.

(b) Noise power.

Figure 5.3: Signal power for different densities. (a): Here, we show the population averaged signal power for both cortex (in red) and thalamus (in blue). The signal power clearly decreases as the stimulus density increases (Krukal-Wallis test; P<0.001). (b): Whilst the signal power clearly shows a density driven decrease, the noise power remains relatively constant.

mean. Thus, given our observations of the raw data, we would predict that signal power should decrease as a function of increasing density.

In figure 5.3 (a), we explicitly calculate the signal power for the responses to the different densities, and average over both cortical and thalamic populations. A decrease can be quite clearly observed, in both populations, indicating a large drop in stimulus-related variability as the stimulus density increases. These changes are statistically significant (Kruskal-Wallis test; P<0.001). This serves to quantify the observations we made in figure 5.2. Interestingly, as the signal power decreases, the noise power stays relatively constant (the difference in noise power as a function of density is not statistically significant). This implies that the amount of variability within the noise is not changing as the density increases, it is purely the amount of variability elicited by the stimulus which shows a decrease.

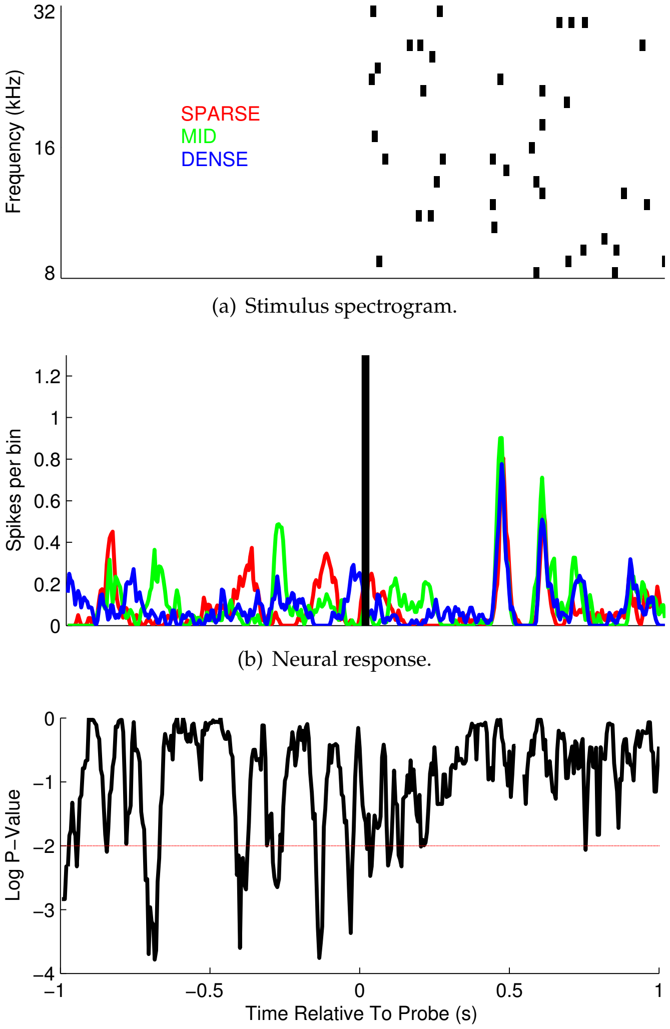
5.3.2 No Contextual Dependence on Stimulus Density

We specifically chose to design the stimulus with a particular switching structure. Rather than allowing the density to switch at random however, we imposed a constraint such that every density had to be preceded by every other density, including its own (this is detailed graphically in figure 5.1 (d)). What this means is that we can explicitly analyse the neural responses to a given density in *context*; that is, we can treat the response to a particular density as a *probe* stimulus, and observe how responses to the probe change when they are preceded by either the same, or a different, density. This particular experimental design was inspired by the work of Asari and Zador (2009), who used a similar paradigm (with a selection of different stimuli) in order to show long lasting contextual dependence in intracellular traces recorded from rat auditory cortex. We wondered if we would be able to observe similar behaviour using our extracellular recordings, when just the spectrotemporal density of a complex stimulus was changed.

Figure 5.4 shows what is essentially a representative example of contextual dependence, the same trend of which is prevalent throughout both cortical and thalamic populations. This particular example focusses on the use of one of the sparse density tokens as a probe stimulus, but similar results are achieved regardless of the identity of the probe. In reference to the particular example shown in figure 5.4, we were interested in establishing if the neural response to this sparse stimulus was affected in some way, by the density of the stimulus which preceded it. Figure 5.4 (a) shows the stimulus spectrogram over a two second period, one second before and after a density transition point (time is denoted relative to probe, so the onset of the sparse stimulus is given at time zero). The preceding stimulus (before time zero) can be either sparse, mid, or dense, and the colour of the lettering indicates the relevant identity of the responses below. Thus, we have three different transitions here; from sparse to sparse, from mid to sparse, and from dense to sparse. In (b) we explicitly show the PSTHs (trial averaged neural responses) to these three density transitions. If we focus purely on the responses to the preceding stimuli, then it seems relatively clear that the different densities evoke somewhat different responses, as one would expect. In contrast to this, the responses to the sparse stimulus, regardless of which density they were preceded by, tend to elicit a stereotyped response, suggesting that the spectrotemporal density of a preceding stimulus does not have a significant effect upon the response to a probe.

This is further quantified in (c), whereby we followed Asari and Zador (2009) and performed a nonparametric Kruskal-Wallis test to assess statistical significance at each point in time. This amounts to performing multiple statistical tests, one for each 5 ms time bin. In each of these 5 ms bins, each response has 20 spike counts associated with it, where 20 is number of trials within the DRC presentation. Thus, the null-hypothesis of the statistical test is simply that the neural responses within the three groups are the same. The resultant P-values are shown in (c). Here, we are showing the log P-values, purely for ease of visualisation. The red line denotes a threshold corresponding to P=0.01.

In the responses to the preceding stimuli, where we would expect the neural responses to be different, we can see extended regions of statistical significance (where



(a) Stimulus spectrogram.

(b) Neural response.

(c) Log P-value.

Figure 5.4: Contextual dependence on stimulus density. Here, we show a representative example of contextual dependence in cortex, the same trend of which is prevalent throughout both cortex and thalamus. (a): Here, we show the stimulus spectrogram over a two second period, one second before and after a density transition point. The stimulus after this transition point (probe) is sparse. The preceding stimulus, can be either sparse, mid, or dense, and the colour of the lettering indicates the relevant identity of the responses below. (b): Neural responses to the stimuli included above. These are all PSTHs, averaged over 20 presentations of DRC. The black line situated at 0 s denotes the transition point to a probe stimulus. Thus, everything that lies on the right of this line is in response to the sparse segment of stimulus shown above. On the left of this line are the responses to the preceding density, be it sparse (red), mid (green), or dense (blue). (c): Statistical significance. We tested for statistical significance between the three groups, using a nonparametric Kruskal-Wallis test. The y-axis here shows log P-values, so that significant values can be readily identified. The red line denotes a threshold of P=0.01. In response to different densities, the neural response is typically different, leading to extended regions of statistical significance. In response to a sparse density, regardless of which density preceded it, the neural response is largely stereotyped, with very few regions of statistically significant difference.

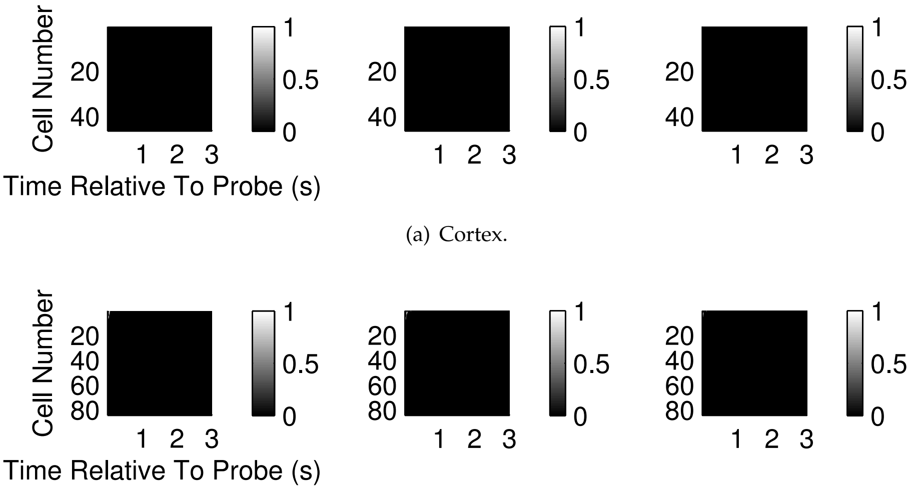
the P-value drops below threshold). These dips correspond directly to regions within the neural responses that are most different. In contrast to this however, there are no extended regions of statistical significance between probe responses (in fact, at only three points in this example did the P-value drop below threshold, which, due to the problem of multiple comparisons, was not enough to yield statistical significance). This is indicative of the probe stimulus eliciting a stereotyped neural response, regardless of the identity of the stimulus that precedes it. Of course, this is purely for one example from the cortical population, and we were particularly curious to establish whether this trend was present throughout the rest of the cortical and thalamic populations. To this end, we computed temporal vectors of P-values (analogous to (c)), for every cell in the population, and every sparse probe stimulus token (of which there are three). Figure 5.5 shows these P-values for both cortex and thalamus (different rows), and each sparse stimulus token (different columns). Note that here, the plotted P-values have been corrected for multiple comparisons. Each row in these plots corresponds to 600 different statistical tests (3 seconds in 5 ms bins) and thus, it likely that at the 0.01 level, some of these results will be due to chance. The P-value has been adjusted to reflect this. As can be observed, there are no regions of statistical significance within cortex. In thalamus, only five cells showed some context dependence, within the first 200 ms after probe onset, but the rest of the population showed nothing. This result was consistent across the use of every density as a probe stimulus (figures not shown).

As a result of this, we were curious whether either linear or nonlinear modelling techniques could be utilised to shed light on any contextual processing that may be taking place, latent within the neural responses.

5.3.3 Modelling Nonlinear Sensitivities to Stimulus Context

5.3.3.1 Changes of STRFs with Sound Density

The changes of linear receptive field estimates with sound density is something that was first detailed by Blake and Merzenich (2002) in the auditory cortex of the owl monkey. In this study they reported systematic changes in receptive field structure as a function of the stimulus environment. These structural changes amounted to both a spectrotemporal sharpening and a decrease in the amount of excitation present within the receptive fields as the stimulus became increasingly dense. Ultimately, they concluded that the auditory system represents a single tone pip with increased specificity, and by fewer action potentials, as the sound density increases. Similar results have also been obtained



(a) Cortex.

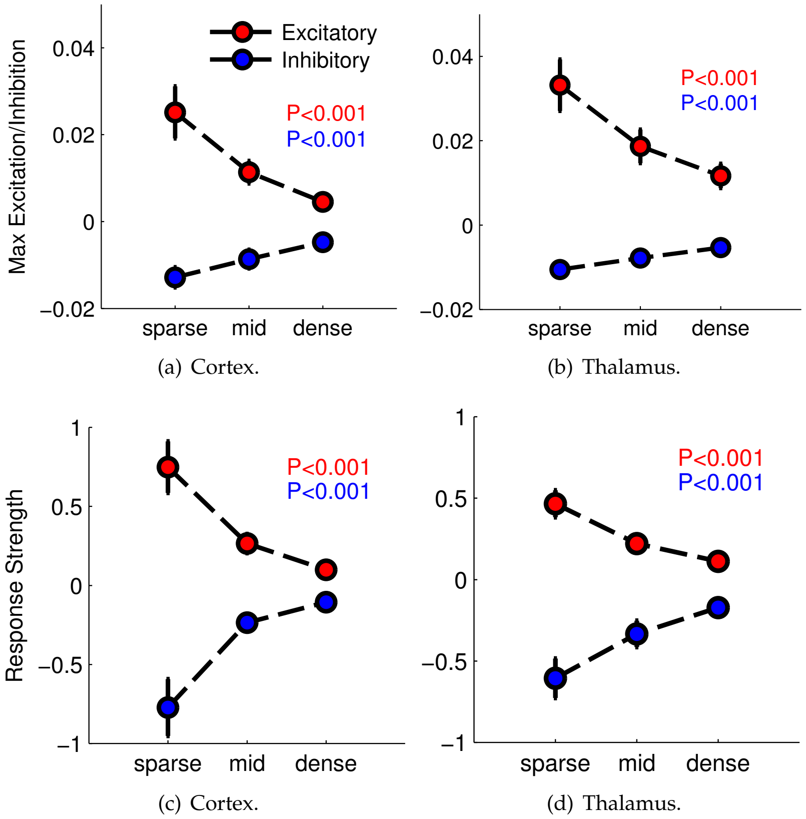
(b) Thalamus.

Figure 5.5: Population context dependence on stimulus density. (a) and (b): Cortex and thalamus. Here, we show P-value rasters from every cell in both populations, sorted by maximum temporal extent of dependence. The probe stimulus here is the sparsest density, and the columns show the three different tokens of this density within the stimulus. P-values have been corrected for multiple comparisons. There is no context dependence within cortex. Five cells within the thalamic population show limited context dependence at ~200 ms after probe onset.

by Valentine and Eggermont (2004); Noreña et al. (2008) in cat.

Our experimental design is very different to what has been previously used to study changes in stimulus density (Blake and Merzenich, 2002; Valentine and Eggermont, 2004; Noreña et al., 2008). In addition to the use of both a different animal model, and different anaesthetic protocol, our DRC stimulus contains different densities. The densest stimuli utilised by Blake and Merzenich (2002) is 1 tone pulse/8 ms/octave. By comparison, our sparsest stimuli is 0.5 tone pulses/5 ms/octave, and our densest stimuli is 2 tone pulses/5 ms/octave. Thus, for the most part, our spectrotemporal densities are denser that what has been utilised in the literature before. Frequency ranges were also different between studies, which has an impact on the overall density of the sound. Given these differences however, we were curious to what extent these previously documented effects were present within our cortical recordings, and how they differ in the thalamus.

For each cell in the cortical and thalamic populations, we fit linear STRF models to the different densities within the switching stimulus (since the stimulus switches every three seconds, we extracted the stimulus segments corresponding to each density, and the corresponding spike times, and concatenated them together). We were ultimately



(a) Cortex.

(b) Thalamus.

(c) Cortex.

(d) Thalamus.

Figure 5.6: Decrease in neural activity with increasing density. *Top row*: Here, we show the population change in maximum (peak) excitation and minimum (trough) inhibition, in both cortex and thalamus. The trends are consistent, and statistically significant (Kruskal-Wallis test, P¡0.001). As stimulus density increases, the amount of excitation and inhibition present within the receptive fields systematically decreases. *Bottom row*: Here, we show the population change in response strength. Again, a systematic and statistically significant decrease in both excitatory and inhibitory response strength can be observed in both areas.

interested in assessing three key things:

- We computed the maximum and minimum values of each STRF, in order to look at how the maximum levels of excitation and inhibition varied across the populations.
- We thresholded the STRFs at half-maximum (for excitation) and half-minimum (for inhibition), and proceeded to sum all of the positive (or negative) time-frequency elements. This gives a measure of excitatory/inhibitory response strength.
- We calculated spectral bandwidths, to establish if any spectral sharpening occurred over the populations.

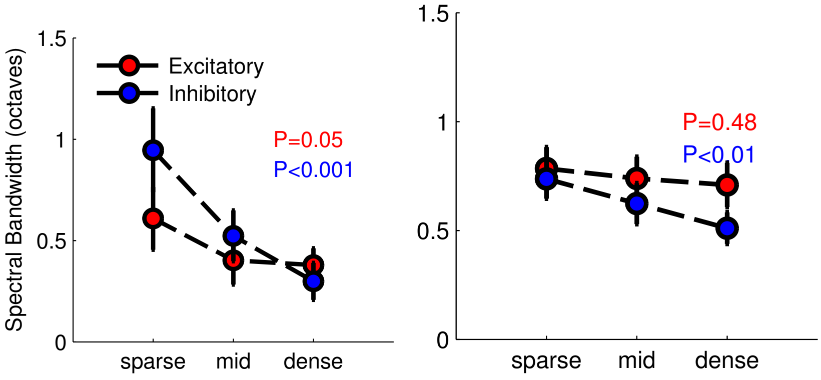
One of the most notable differences in the STRFs, is that as the stimulus density increases, the amount of excitation and inhibition within the receptive field drops quite dramatically. This result is consistent across both cortical and thalamic populations, and is detailed in figure 5.6. We quantified this decrease in terms of both maximum (peak) excitation and minimum (trough) inhibition, and also excitatory and inhibitory response strength.

Both of these measures yield the same, statistically significant, trends indicative of a higher stimulus density eliciting a weaker neural response (shown in figure 5.6). One interesting point to note however, is that even though the decrease in max excitation/inhibition is statistically significant, the excitatory drop (from ~0.025 to ~0.005) is larger than the corresponding decrease in inhibition (from ~-0.01 to ~-0.005). This trend is visible in both cortex and thalamus (figure 5.6 (a) and (b)).

We also assessed the extent to which spectral sharpening was present across both populations (assessed by spectral width at half maximum (for excitatory bandwidths) or half minimum (for inhibitory bandwidths)). In the cortical population, we observed a decrease in excitatory spectral bandwidth between the sparse and mid densities (figure 5.7), consistent with the idea that sounds in dense acoustic environments are represented with increased specificity in cortex. We did not observe a corresponding drop in bandwidth between the mid and dense stimuli however. We also observed a significant decrease in inhibitory spectral bandwidth, something that has not been reported previously.

In the thalamic STRF population, bandwidths were more similar over the different densities. We observed a significant change in inhibitory spectral bandwidth but the excitatory spectral bandwidth remained relatively constant (figure 5.7 (b)).

It is notable that the primary difference in our observations with what has been reported previously (Blake and Merzenich, 2002), is a reduction in the amount of inhibition present within the receptive fields. Aside from the use of different densities within the stimulus, one of the fundamental differences with our study is our rigorous use of regularisation whilst estimating the statistical models (this was treated in chapter 3). Such regularisation was not utilised in any of the previous STRF density studies. Due to the fact that we observe such a significant reduction in stimulus-related variability as the density increases, it is likely that the regularisation will act to “shrink” the parameters somewhat, due to a poorer model fit. This could potentially account for some of the additional differences between studies that we observe.



(a) Cortex.

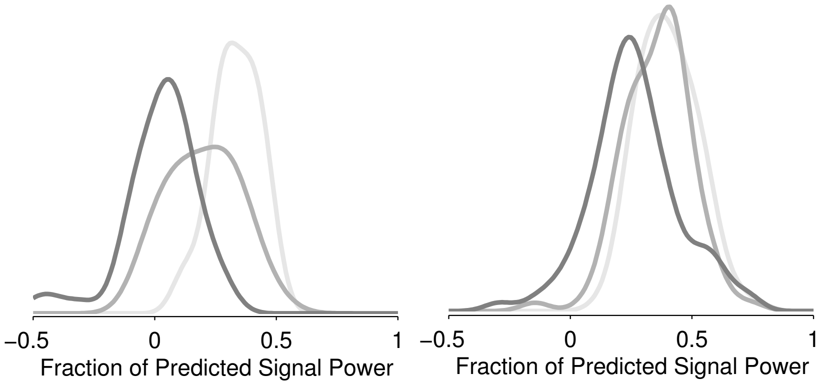
(b) Thalamus.

Figure 5.7: Changes in spectral bandwidth. Spectral bandwidth was assessed by calculating the spectral width at either half maximum (for excitation) or half minimum (for inhibition). In cortex, a decrease in excitatory bandwidth can be observed between the sparse and mid stimuli, in addition to a systematic decrease in inhibitory bandwidth. The decrease in excitatory spectral bandwidth over the range of densities in cortex is not statistically significant however (Kruskal-Wallis; P=0.05), but the decrease in inhibitory spectral bandwidth is (Kruskal-Wallis; P<0.001). In thalamus, the excitatory bandwidth remains relatively constant (Kruskal-Wallis; P=0.48), whilst the inhibitory bandwidth decreases (Kruskal-Wallis; P<0.01).

5.3.3.2 Predictive Capability of Linear and Multilinear Models

Since we actually had a measure of the stimulus-related variability for each of the cells in response to the different stimulus densities, we were able use this information to evaluate the predictive capability of the learnt STRFs. As in the previous chapter (section 4.2.7, we used a definition of predictive power, which is essentially a measure of explainable variance, but whereby we use the signal power to give us an estimate of how much stimulus-related variability can be explained by a given model.

In cortex, the fraction of signal power successfully predicted by the linear STRF models clearly decreased as the stimulus density increased. This is shown in figure 5.8 (a), where density estimates of the STRF predictive powers have been provided. Linear predictions of responses to the sparse stimulus in cortex yielded a cross-validated extrapolated predictive power of around 34%. This number decreases to 28% and then 15% as the stimulus becomes increasingly dense. This is evident in the leftward shift of the density estimates in figure (a). Interestingly, it can also be observed that, at higher densities, there are a far larger number of predictive powers that fall below zero, indicating a poor model fit. In the thalamus, this trend is similar, but not as pronounced (in (b)). Noticeably, the thalamic responses were easier to predict using a linear model,



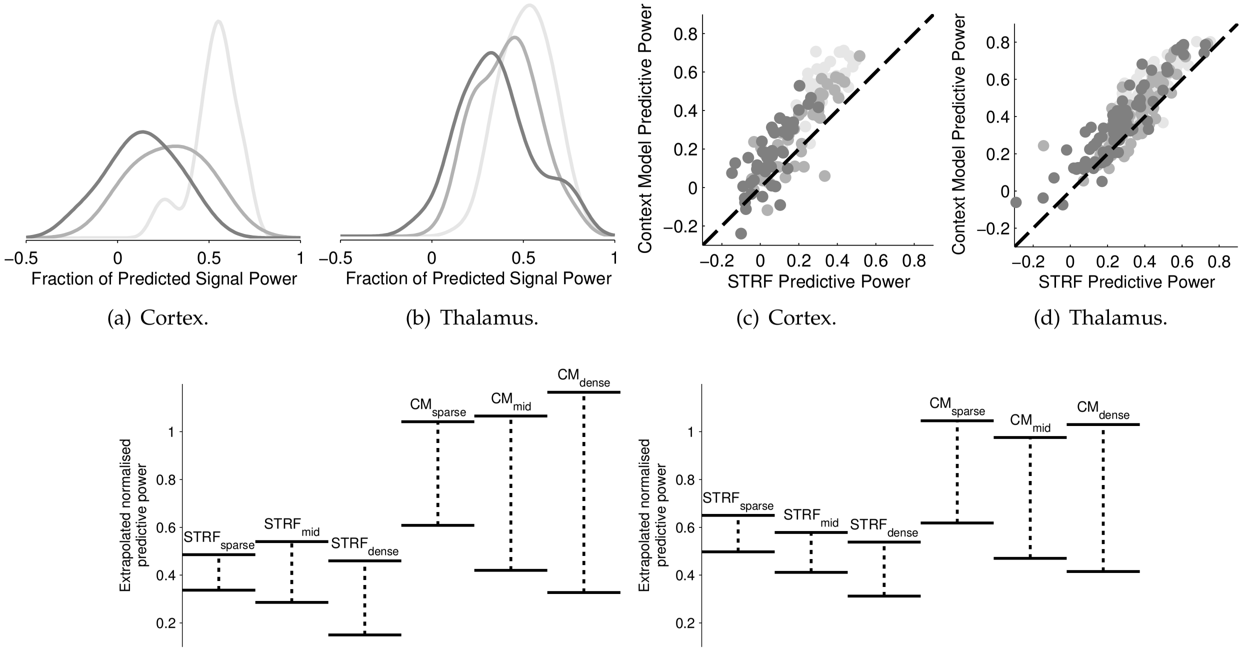
(a) Cortex.

(b) Thalamus.

Figure 5.8: Response nonlinearity increases with stimulus density. Here, density estimates of the STRF predictive power distributions in both cortex (left) and thalamus (right) have been provided. The leftward shift that is particularly evident in cortex is indicative of encoding nonlinearities playing a more important role in particularly dense acoustic environments.

indicating (as in the previous chapter) that thalamic responses are somewhat more linear than cortex. The sparse stimulus yielded a cross-validated predictive power of 50%, which dropped to 42% and then 40%, as the density increased. In (b), this leftward shift can be observed.

Such a result has implications for how we can think about the role of nonlinearities in diverse acoustic environments. The fact that a linear model's predictive capability decreases as the stimulus density increases is indicative of the fact that such nonlinearities play a more significant role as an acoustic stimulus becomes denser and more complex. We were particularly curious as to how much of a role stimulus context played in shaping the neural responses to what are essentially stimuli of increasing complexity. To this end, we fit multilinear context models (as described in the previous chapter) to each cell within the population.



(a) Cortex.

(b) Thalamus.

(c) Cortex.

(d) Thalamus.

(e) Cortex.

(f) Thalamus.

Figure 5.9: Context model predictive capability. (a) and (b): Density estimates of the context model predictive power distributions in both cortex and thalamus (the corresponding STRF density estimates were shown in figure 5.8. (c) and (d): Scatter plots showing the increase in predictive power that the context model achieves over the corresponding STRF estimates. (e) and (f): Extrapolated predictive powers for all densities, and both cortex and thalamus. The upper part of the bars shows the training-set predictive power, whilst the lower part of the bars shows the cross-validated test-set predictive power.

Before we proceed to discuss the structure observed within the context model, we will first summarise the predictive benefit that it yields. Ultimately, the context model provided a substantial gain in predictive ability over a simple linear model at all densities, and for both cortex and thalamus. Figure 5.9 summarises several key elements of this analysis. In (a) and (b), the predictive power density estimates of the context model have been shown. Although the predictive power decreases as a function of density (as does the linear model), the predictive increase of the context model over the corresponding STRF estimates is particularly noticeable ((c) and (d)).

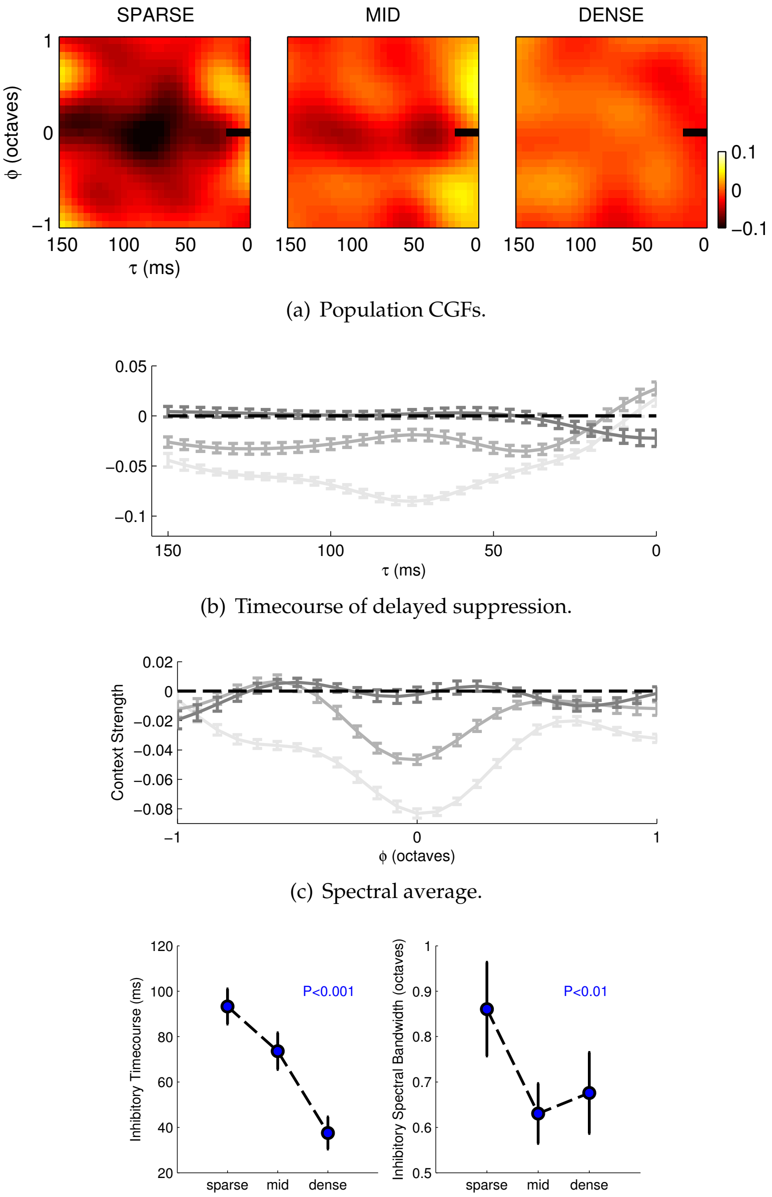
A linear model provided a particularly poor fit to the cortical data, especially at higher densities (as was evident from the large number of negative predictive powers in 5.9 (a)). The context model has improved on this dramatically, providing quite a substantial gain. In the sparse case, the lower bound has risen from 34% to 61%, almost doubling the predictive power of the model. Similar gains can also be observed the higher densities with mid rising from 28% to 42%, and dense rising from 15% to 32%. These are summarised in figure 5.9 (e).

The thalamus also saw substantial gain through the use of the context model, although this gain was not as large as seen in cortex. In the sparse case, the rise in predictive power was from 50% to 63%, in mid, from 42% to 48%, from 40% to 53%. Interestingly, even though the lower bound extrapolation does yield a larger predictive power in the dense case than the mid case, the actual density estimates of the predictive power distribution (shown in figure 5.9 (b)) still clearly shows a downward shift as the density increases. It may be that the slight deviation in extrapolated value is due to the linear fit being somewhat biased by outlier values. As before, these statistics are summarised in figure 5.9 (f).

5.3.3.3 CGF Structure in Cortex and Thalamus

Given that the context model provided a substantial gain in predictive power over a simple linear model, it is quite clear that stimulus context must play an important role in shaping the neural response in complex sound environments. As a result of this, we proceeded to analyse the structure of the learnt context models, in an attempt to establish the nature of the nonlinear interactions at work.

In the previous chapter, we showed that such a context model, when fit to both cortical and thalamic responses to DRC stimuli, yields CGFs that contain particularly rich spectrotemporal structure. This structure becomes particularly apparent when popula-



(a) Population CGFs.

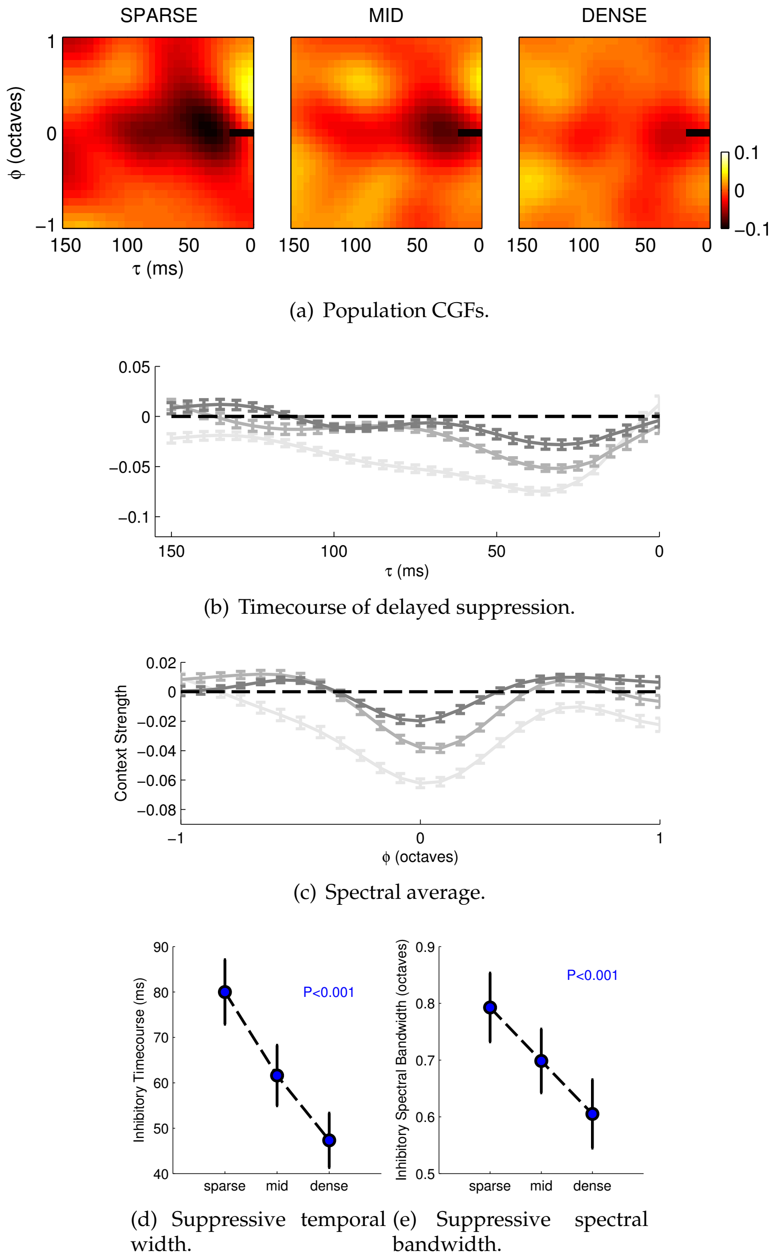
(b) Timecourse of delayed suppression.

(c) Spectral average.

(d) Suppressive temporal width.

(e) Suppressive spectral bandwidth.

Figure 5.10: Population CGF structure in cortex. This plot summarises the CGF analysis that was carried out in cortex. (a): Population CGFs. (b): The average timecourse of delayed suppression for each density. Here we have averaged across an 1 octave window surrounding the current tone. In doing this, we can observe the minimum of the suppressive trough shifting to the right as the stimulus density increases. (c): Spectral average. Here, we can see the magnitude of the weights changing, and more importantly a decrease in spectral bandwidth (in the sparse and mid cases). (d) and (e): Suppressive temporal width and suppressive spectral bandwidth. There is a systematic decrease in the suppressive temporal width at all densities, and a decrease in spectral bandwidth between the sparse and mid stimulus (the dense stimulus does not contain much average spectral structure in cortex).



(a) Population CGFs.

(b) Timecourse of delayed suppression.

(c) Spectral average.

(d) Suppressive temporal width.

(e) Suppressive spectral bandwidth.

Figure 5.11: Population CGF structure in thalamus. This plot summarises the CGF analysis that was carried out in thalamus. (a): Population CGFs. (b): The average timecourse of delayed suppression for each density. Here we have averaged across an 1 octave window surrounding the current tone. In doing this, we can observe the minimum of the suppressive trough shifting to the right as the stimulus density increases. (c): Spectral average. Here, we can see the magnitude of the weights changing, and more importantly a decrease in spectral bandwidth. (d) and (e): Suppressive temporal width and suppressive spectral bandwidth. There is a systematic decrease in both the suppressive temporal width and spectral bandwidth at all densities.

tion averages are considered, and largely consists of a region of delayed suppression, and a near-simultaneous region of enhancement.

The CGF structure that we observe in the current data changes in a systematic way as the stimulus density increases. The results of this CGF analysis are shown in figure 5.10 and 5.11. In the sparse CGFs within both cortex and thalamus, the primary discernible feature seems to be a large region of delayed suppression. The temporal extent of this region is similar to what we reported in the previous chapter. Here, the temporal extent in cortex is the full 150 ms range of the CGF. In thalamus, the delayed suppression does still last for the full duration of the CGF (150 ms), but significantly decreases in efficacy after around 100 ms. Interestingly, although the full temporal extent of the suppression is somewhat lengthy, the minimum suppression is reached after only 75 ms in cortex, and 40 ms in thalamus. These statistics are highlighted in 5.10 and 5.11.

In addition to this delayed region, there is also a region of simultaneous enhancement, present in both areas. In this particular version of the stimulus however, the temporal resolution has been increased to 5ms, due to the jittered nature of the tones within the spectrogram. In the reports of simultaneous enhancement in the previous chapter, the resolution was fixed at the lengthier 20 ms and thus simultaneous in that context was sound energy occurring simultaneously within a 0-20 ms bin. The finer resolution here allows greater specificty.

In cortex, this simultaneous enhancement is situated at either side of the current tone, providing an enhancement in gain whenever sound energy is present above or below the frequency a given tone. This gain seems to be somewhat asymmetric in this particular dataset however, with the temporal extent of the enhancement lasting around 20 ms on the negative side, and almost 40 ms on the positive side. Given the resolution of the simultaneity that we reported in the previous chapter, these timescales are still consistent. In thalamus, this simultaneous enhancement occurs only on the positive side of the current tone, and lasts for around 25 ms.

As the density of the stimulus increases however, there seems to be one primary systematic change in the CGFs that occurs in both cortex and thalamus. The spectrotemporal range of the delayed nonlinear interactions shrinks as the stimulus becomes denser and more complex. This is shown in 5.10 (b), and (c), and 5.11 (b), and (c). In both cases, (b) and (c) show temporal averages over a 1 octave window around the current tone. This shows the suppressive trough shifting to the right as a function of density, explicitly shortening the temporal extent of the nonlinear interactions. This is particularly evident in cortex, where the minimum of the suppressive region moves from 80 ms in the sparse case, to 45 ms in the mid case, and only 10 ms in the dense case.

In thalamus, these timescales are somewhat shorter, with the minimum in the sparse CGF at around 40 ms, and then shifting by approximately 5 ms each time the density increases. Figures (c) and (d) show spectral averages across the entire CGF. Here, we can see the clear spectral sharpening that occurs as the density of the stimulus increases. In order to quantify this further, we explicitly calculated both the suppressive temporal width and suppressive spectral bandwidth within the different populations (by calculating the relevant width at half minimum of either a temporal or spectral strip through the trough in the CGF). These results are shown in figures (d) and (e). In both cortex and thalamus, the suppressive temporal width shows a statistically significant, systematic decrease as a function of density (which is clearly observable from the structure present within the population CGFs). In addition, the suppressive spectral bandwidth decreases systematically within the thalamus. In cortex, the bandwidth decreases significantly between the sparse and mid stimulus, but the change between mid and dense is not significant (this is also relatively clear in (c), since the dense cortical CGF does not contain much average spectral structure).

In addition to the delayed suppression that persists throughout the CGFs, the simultaneous enhancement also appears at higher densities. This is evident particularly in the mid CGFs, and also, to a somewhat limited extent, in the dense CGFs.

As in the previous chapter, we were curious as to whether or not any potential differences between the PRF component of the context model, and the STRF population, could help to elucidate what could potentially be going on. A table of such spectrotemporal profile differences is shown in 5.1. We were particularly interested in the relative amount of suppression contained within each population. If, as we saw in the previous chapter, the STRF population contained more relative suppression than that of the PRF population, we could potentially attribute this suppression to the nonlinear effects of stimulus context.

For all densities and in both cortex and thalamus, the STRF population contains a greater amount of relative suppression than the corresponding PRF population. This is particularly interesting due to the rich suppressive structure that is present within the CGFs. The fact that there is less suppression within the PRF population is indicative of the fact that the context model has been able to attribute some of the STRF suppression to the nonlinear effects of stimulus context. As an example of this, some of the inhibitory

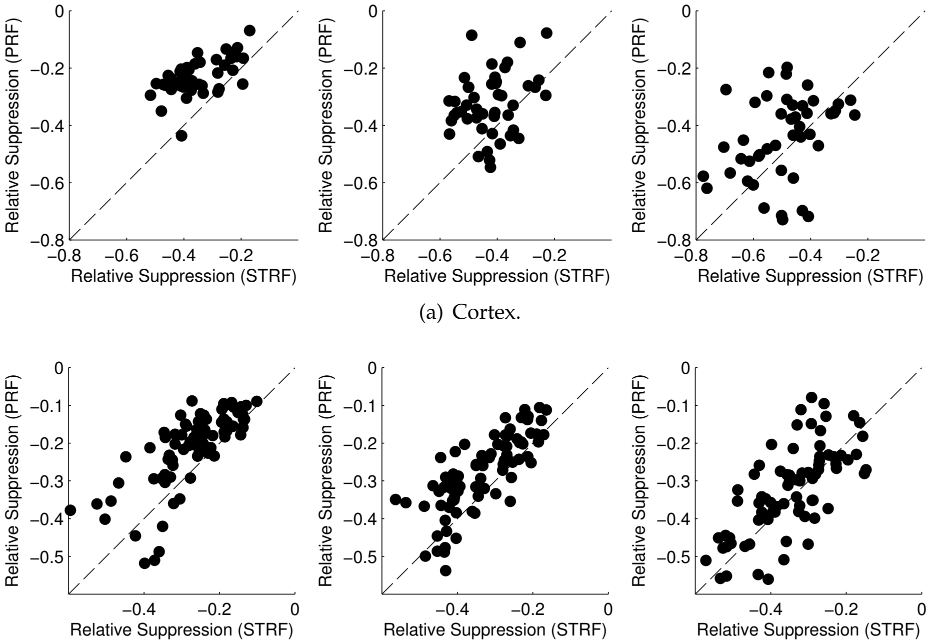
CORTEX	LOW	MID	DENSE
STRF Receptive Field Duration (ms)	113.28 ± 1.39	112.59 ± 0.89	99.83 ± 1.87
STRF Excitatory Subfield Width (ms)	37.07 ± 0.79	41.38 ± 0.66	36.72 ± 1.25
STRF Inhibitory Subfield Width (ms)	48.45 ± 1.72	41.03 ± 1.30	51.03 ± 1.41
STRF Excitatory Bandwidth (octaves)	0.61 ± 0.03	0.40 ± 0.02	0.38 ± 0.02
STRF Inhibitory Bandwidth (octaves)	0.95 ± 0.04	0.52 ± 0.02	0.30 ± 0.01
PRF Receptive Field Duration (ms)	94.48 ± 2.14	106.38 ± 1.42	99.48 ± 1.86
PRF Excitatory Subfield Width (ms)	39.66 ± 1.06	46.90 ± 13.34	52.07 ± 2.07
PRF Inhibitory Subfield Width (ms)	49.31 ± 1.49	43.28 ± 1.38	58.10 ± 1.58
PRF Excitatory Bandwidth (octaves)	0.65 ± 0.02	0.54 ± 0.02	0.41 ± 0.02
PRF Inhibitory Bandwidth (octaves)	0.54 ± 0.02	0.40 ± 0.02	0.34 ± 0.01
THALAMUS			
STRF Receptive Field Duration (ms)	102.22 ± 0.69	93.53 ± 0.87	87.47 ± 1.01
STRF Excitatory Subfield Width (ms)	18.86 ± 0.22	17.09 ± 0.23	15.06 ± 0.27
STRF Inhibitory Subfield Width (ms)	48.92 ± 0.64	45.70 ± 0.55	42.97 ± 0.59
STRF Excitatory Bandwidth (octaves)	0.78 ± 0.01	0.74 ± 0.01	0.71 ± 0.01
STRF Inhibitory Bandwidth (octaves)	0.74 ± 0.01	0.62 ± 0.01	0.52 ± 0.01
PRF Receptive Field Duration (ms)	81.84 ± 0.96	96.20 ± 0.78	83.92 ± 0.93
PRF Excitatory Subfield Width (ms)	19.68 ± 0.23	21.14 ± 0.33	19.49 ± 0.39
PRF Inhibitory Subfield Width (ms)	44.49 ± 0.59	49.30 ± 0.60	49.49 ± 0.70
PRF Excitatory Bandwidth (octaves)	0.88 ± 0.01	0.683 ± 0.01	0.68 ± 0.01
PRF Inhibitory Bandwidth (octaves)	0.70 ± 0.01	0.61 ± 0.01	0.52 ± 0.01

Table 5.1: Summary of spectrotemporal profile differences between stimulus densities.

profiles (subfield width and spectral bandwidth) within the STRF population show systematic changes as the density increases. The inhibitory spectral bandwidth decreases as a function of density in both cortex and thalamus, and the inhibitory subfield width also shows a decrease with increasing density (although this change is largely isolated to the transition from sparse to mid). The fact that we such a clear spectral and temporal sharpening within the delayed suppressive region of the CGFs indicates that some of the observable changes within the STRF population could be attributable to the nonlinear effects of stimulus context.

5.3.3.4 CGF Structure Under Alternative Anaesthesia

Prior to carrying out the experiments that yielded the primary data that has been discussed in this chapter, we carried out a set of control experiments (largely to verify the efficacy of the stimulus). These control experiments were only carried out in the auditory thalamus, and utilised an identical stimulus to what has been described previously. The primary difference however, is that these control experiments were carried out utilising urethane anesthesia (as opposed to the ketamine/medetomidine protocol described earlier).



(a) Cortex.

(b) Thalamus.

Figure 5.12: Density elicited relative suppression. We directly compared the amount of relative suppression (given by min(filter)/max(filter) − min(filter)) in both cortex and thalamus, for all stimulus densities (sparse - dense, left -right). The trend in results is uniform, whereby for the majority of cells, the amount of relative suppression is greater within the STRF than the corresponding PRF. This suggests that the CGF component of the context model is better able to account for the suppressive effects of stimulus context.

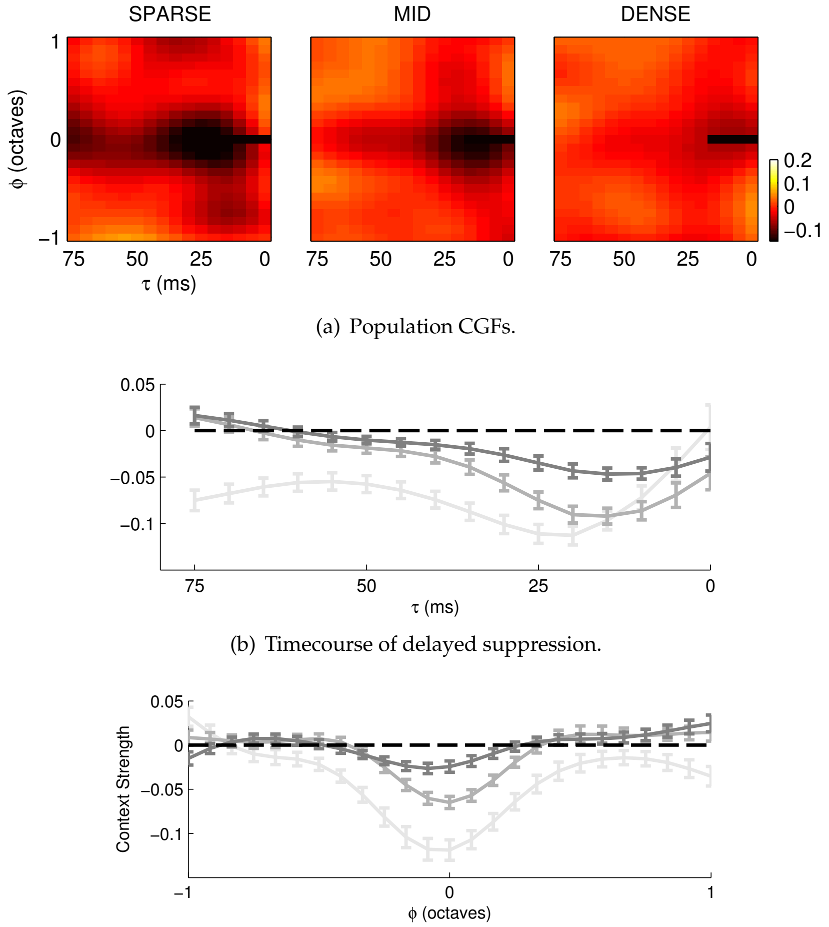
This section merely exists to point out that all of the results presented thus far, that have been presented utilising the ketamine data, also hold within the control dataset. The CGF structure is shown in figure 5.13 for comparison.

5.4 Discussion

Our main finding throughout the course of this study is that auditory cortical and thalamic processing involves significant nonlinear contextual interactions in both sparse and dense acoustic environments.

5.4.1 Response Reliability

We first showed that neural responses in both cortex and thalamus were more reliable in sparse stimulus environments. As the stimulus density increased and became in-



(a) Population CGFs.

(b) Timecourse of delayed suppression.

(c) Spectral average.

Figure 5.13: Population CGF structure in thalamus using alternate anaesthesia. Here, we present the CGF structure for a set of control experiments that were carried out in the auditory thalamus using urethane anaesthesia. For the most part, the structure is incredibly similar to what has been observed using the ketamine/medetomidine protocol. Both the structure present within the population CGFs, and the timescales of the delayed suppressive regions are similar to what was shown in figure 5.11.

creasingly complex, this reliability dropped, resulting in a decrease in stimulus related variability. We were able to quantify this change using the statistical tools developed by Sahani and Linden (2003b). This obviously raises the question, why does the stimulus-related variability decrease with stimulus density? One explanation is simply due to the density of our sparsest stimulus component, averaging 0.5 tone pulses per bin per octave. This means that there are numerous periods within the spectrogram where very few (or even no) tones are playing. As a result of this, almost everything within the sparse component of the DRC stimulus seems like an onset. It is well known that a vast number of auditory neurons throughout the auditory pathway respond well to such transients (Heil (1997a,b), specifically, in cortex (Eggermont, 1993), and in thalamus (Rouiller et al., 1981)). In fact, figure 5.4 shows such an example of this. Here, three separate responses to such a sparse segment is shown. The responses themselves have clear transient peaks (regions of high variability) which correspond directly to the onsets present within the spectrogram plotted above. As the density increases, such onsets becomes less common, and the stimulus-related variability decreases as a result.

5.4.2 Contextual Dependence

We were particularly enthusiastic about the inherent switching nature that we had placed on the stimulus, such that we could study the effects of spectrotemporal density in *context*. This was inspired by the paradigm of Asari and Zador (2009), who studied the problem of context dependence in the rat auditory cortex, using intracellular means. Here, they utilised a number of different natural and synthetic sounds, and were able to show long lasting context dependence (of up to 4 s) within their subthreshold responses. One immediate difference between our study and theirs (aside from the obvious intracelluar/extracellular difference) is the choice of stimuli. Here, we used one particular class of stimuli (the DRC), and systematically varied the parameter which controls spectrotemporal density. In Asari and Zador (2009), multiple stimulus types were used which could, in principle, lead to a difference in contextual dependence.

When dealing with extracellular recordings, at the resolution of 5 ms, statistical analysis becomes difficult. One of the major problems that we had was that, for every time bin, a significance test had to be carried out. Given the fact that some of the densities elicit such poor responses, in a small time window, this can lead to performing significance tests between vectors consisting of mainly zero elements. This is not ideal. As a result of this, we tried several approaches, including simply increasing the bin size, and also smoothing the data in increasingly elaborate ways. None of these approaches were able to yield statistical significance.

5.4.3 Nonlinear Sensitivities to Stimulus Context of Increasing Complexity

Given that we were not able to see any significant amount of contextual dependence, we proceeded to perform linear and multilinear analyses, to potentially identify nonlinear contextual interactions, latent within the neural response.

We observed a significant decrease in the maximum excitation within the STRF population, and the overall excitatory response strength, as the stimulus density increased. This decrease in neural responsiveness corroborates previous studies in this area (Blake and Merzenich, 2002; Valentine and Eggermont, 2004; Noreña et al., 2008). Furthermore, we extended the same analysis to STRFs within the auditory thalamus, and observed the same statistically significant trends.

In cortex, we also showed that the shape of the STRF changed as a function of density, in terms of both excitatory and inhibitory spectral bandwidth. Such a change in excitatory bandwidth was also reported by Blake and Merzenich (2002). The decrease in inhibitory bandwidth is something that has not been reported before, but, as discussed earlier, this may be due to differences in statistical regularisation. In contrast to the spectral sharpening observed in cortex, we did not see a change in excitatory bandwidth within the thalamic population.

Blake and Merzenich (2002) attributed the changes that they observed to synaptic depression. They speculated that if synaptic depression is variable across the different inputs that compose the receptive field, then an increase in density should change the relative contributions of those inputs, which will have the effect of altering the shape of the STRF. Such a claim was further corroborated by David et al. (2009), who illustrated that the stimulus dependence of spectrotemporal tuning can be explained by a model in which the synaptic inputs to cortical neurons are susceptible to rapid nonlinear depression.

In their intracellular study, Wehr and Zador (2005) established that synaptic depression at the thalamocortical synapse was likely the cause of the long-lasting suppressive effects observed in cortical neurons. Moreover, they also showed that such cortical suppression was unlikely to be inherited from thalamic response properties, due to their quick recovery from suppression. This is perhaps indicative that the amount of synaptic depression that the thalamus itself receives as input (presumably from the inferior colliculus), is not significant enough to instigate substantial changes in receptive field structure. This hypothesis corroborates the fact that we do not observe any spectral sharpening within the thalamic population.

Ultimately, the core finding of this study, is that we can successfully use the multilinear framework to estimate nonlinear contextual effects within the neural response, that are not able to be identified using simple linear methods. The structure that we observed within the CGFs at multiple densities in both cortex and thalamus is particularly rich and informative, and the fact that we again observe that the STRF population contains more relative suppression that the PRF population, for all densities, adds credence to the fact that the context model is capable of providing a better way to model suppressive effects.

In the CGFs, we see a systematic decrease in both the suppressive temporal width and suppressive spectral bandwidth as the stimulus density is increased. This systematic decrease occurs in both the cortical and thalamic populations, albeit on a far faster timescale in thalamus. This provides insight into temporal integration for different acoustic conditions. Clearly, in a sparse environment, where there are very few acoustic inputs, the auditory system can allow for longer temporal, and broader spectral, integration. This allows for incoming acoustic information to be integrated over a larger spectrotemporal window. As the stimulus environment becomes increasingly complex however, in order to adequately integrate information, it makes intuitive sense to shorten this window and become more spectrotemporally selective.

The fact that the STRF itself exhibits structural changes related to density has important behavioural consequences. In a dense, noisy environment, the neural response is more spectrotemporally selective, and thus more capable of representing auditory "edges", which are highly prevalent in narrowband sounds. In contrast to this, in a sparse, quiet environment, the neural response is far less selective, and will respond to broadband features in both frequency and time. The context model provides a potentially mechanistic viewpoint for understanding such processes, and our results suggest that they can be attributed to nonlinear changes in stimulus context, or interpreted as a form of contextual gain control, whereby the amount of influence an environment has on the neural response is modulated based on its own acoustic context.

VI

Conclusions

6.1 Stimulus-Response Functions in Audition

Understanding how complex sounds elicit neural responses is a particularly important goal within auditory neuroscience. It is because of this that the use of neural encoding models has become especially popular in recent decades. Within the auditory-research community especially, the STRF model has been used extensively to characterise auditory function in a number of different brain areas.

Stimulus-response functions are always a simplification of reality. Dealing with a function that simply maps a stimulus to a response is necessarily a "black-box" approach. As a result of this, if such a function can be tailored using knowledge about the underlying biological system, then it may be particularly useful. This is the general ethos of the modelling framework that we have attempted to provide throughout this thesis.

As discussed on several occasions, one of the difficulties with using an STRF model to describe auditory function is that it is inherently linear and neural responses are not. Unfortunately, some of the standard approaches to nonlinear modelling quickly succumb to the curse of dimensionality, requiring large amounts of data to adequately estimate. An example of this is the Volterra series expansion, where the amount of data required increases exponentially with the model order.

The modelling framework used throughout this thesis is an attempt to combine the best of both worlds. The particular parameterisation that we utilise is inspired by biology, and the knowledge that neural responses can be modulated by their acoustic context. Such a parametrisation is not only plausible but it defines explicit structural constraints on interactions within the stimulus, which partially alleviates the need for huge swathes of data in order to adequately estimate.

One of the core results, prevalent throughout, is that the model reveals that there are indeed significantly predictive nonlinear interactions present within neural responses to complex sounds. Chapter 4 illustrates this using data recorded from both auditory cortex and thalamus. In addition to the successful identification of nonlinear interactions within both populations of cells, by applying the model to different fields within the auditory cortex, and different subdivisions within the thalamus, insights can be gleaned as to the nonlinear processing characteristics of these different areas. Moreover, chapter 5 illustrates how such a model can also be used to probe the nonlinear mechanisms that underlie the ability of the auditory system to operate in diverse acoustic environments. Such a framework provides a novel extension to the study of receptive fields in multiple brain areas, and extends existing understanding of the way in which stimulus context drives complex auditory responses.

6.2 Insights into Auditory Function

One of the fundamental results to come from this thesis is the structure that is present within the context model when it is fit to both cortical and thalamic data. This structure is inseparable, and indicates that auditory responses are sensitive to specific spectrotemporal combinations of sound energy with a complex sound stimulus.

The contextual interactions that the model estimates can be thought of as a form of contextual gain control, whereby suppression acts to reduce the effect of both excitation and inhibition, and facilitatory effects act to enhance excitation and inhibition. The form of the interactions that we typically observe (that are especially prevalent at the population level) consist of a lengthy region of delayed suppression, and a near-simultaneous region of enhancement. Such delayed suppression is likely to underlie the temporal analysis of sound, and may be related to known neural mechanisms such as forward suppression, which has been discussed at length in the literature (although the gain control mechanism here is somewhat more general, in that it also allows for a decrease in inhibition, as well as excitation). The simultaneous enhancement seems more likely to reflect the spectral analysis of sound.

Such contextual gain control could also have direct behavioural relevance, and may play an important role in auditory scene analysis. The environment around us is acoustically diverse, consisting of both temporal and spectral complexities. The temporal duration, or spectral extent of such complexities can lead to different acoustic events being perceived either in a single auditory stream, or being segregated into multiple. The neural mechanisms underlying such stream segregation remains elusive. It may be the case that the simultaneous enhancement that we observe is responsible for mediating onset responses, and the grouping of acoustic events that occur at different frequencies with a common temporal onset. The delayed suppression could perhaps mediate the temporal effects of stream segregation, by reducing the efficacy of acoustic events over a short window, thus enabling such events to stay within the same auditory stream.

The heterogeneity of our results is also consistent with previously reported studies of nonlinear sensitivities in cortical neurons (Sadagopan and Wang, 2009). Even though we observe structure on a population level, single cell model fits still show distinct spectrotemporal sensitivities. Thus, the context model seems capable of capturing global nonlinear effects, that likely represent some underlying neural mechanism, present on a grand scale, whilst still retaining the ability to capture individual neural sensitivities to complex sounds.

6.3 Future Applications

The scope of such a multilinear framework is immense. Although this thesis has focussed on characterising responses within the auditory system, the framework itself is applicable to any sensory area. Indeed, it would be fascinating to apply such a model to the visual system, where the effects of stimulus context have been extensively studied in the form of extra-classical receptive fields.

Another important direction would be the use of such a model in behaviour. An example of a particularly relevant paradigm is the work of Fritz et al. (2003), where an animal (in this case, a ferret) is actively engaged in a behavioural task, typically involving the detection of a tone, or a simple discrimination. Over the course of such an experiment, receptive fields are typically estimated in "pre-behaviour", "during-behaviour", and "post-behaviour" conditions. Indeed, it has been shown that the structure within the receptive fields differs to reflect the task that the animal is engaged in. This kind of adaptive, task-related plasticity is something that would be particularly amenable to a multilinear analysis, in order to identify potential nonlinear mechanisms that may shape behaviour.

It is our hope that in the decades to come, such a multilinear approach to neural characterisation becomes commonplace within the sensory community, and is utilised to the same extent that the humble STRF model has been, over the years.

BIBLIOGRAPHY

Aertsen, A. M. and Johannesma, P. I. (1980). Spectro-temporal receptive fields of auditory neurons in the grassfrog I. Characterization of tonal and natural stimuli. *Biological Cybernetics*, 38(4):223–234.

Aertsen, A. M. and Johannesma, P. I. (1981). A comparison of the spectro-temporal sensitivity of auditory neurons to tonal and natural stimuli. *Biological Cybernetics*, 42(2):145–156.

Aertsen, A. M., Johannesma, P. I., and Hermes, D. J. (1980). Spectro-temporal receptive fields of auditory neurons in the grassfrog II. Analysis of the stimulus-event relation for tonal stimuli. *Biological Cybernetics*, 38(4):235–248.

Aertsen, A. M., Olders, J. H., and Johannesma, P. I. (1981). Spectro-temporal receptive fields of auditory neurons in the grassfrog III. Analysis of the stimulus-event relation for natural stimuli. *Biological Cybernetics*, 39(3):195–209.

Ahrens, M. B., Linden, J. F., and Sahani, M. (2008a). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *Journal of Neuroscience*, 28(8):1929–1942.

Ahrens, M. B., Paninski, L., and Sahani, M. (2008b). Inferring input nonlinearities in neural encoding models. *Network: Computation in Neural Systems*, 19(1):35–67.

Aitkin, L. M., Pettigrew, J. D., Calford, M. B., Phillips, S. C., and Wise, L. Z. (1985). Representation of stimulus azimuth by low-frequency neurons in inferior colliculus of the cat. *Journal of Neurophysiology*, 53(1):43–59.

Aitkin, L. M. and Phillips, S. C. (1984). Is the inferior colliculus an obligatory relay in the cat auditory system? *Neuroscience Letters*, 44(3):259–264.

Andersen, R. A., Roth, G. L., Aitkin, L. M., and Merzenich, M. M. (1980). The efferent projections of the central nucleus and the pericentral nucleus of the inferior colliculus in the cat. *Journal of Comparative Neurology*, 194(3):649–662.

Anderson, L. A., Christianson, G. B., and Linden, J. F. (2009a). Mouse auditory cortex differs from visual and somatosensory cortices in the laminar distribution of cytochrome oxidase and acetylcholinesterase. *Brain Research*, 1252:130–142.

Anderson, L. A., Christianson, G. B., and Linden, J. F. (2009b). Stimulus-specific adaptation occurs in the auditory thalamus. *Journal of Neuroscience*, 29(22):7359–7363.

Anderson, L. A. and Linden, J. F. (2011). Physiological differences between histologically defined subdivisions in the mouse auditory thalamus. *Hearing Research*, 274(1-2):48–60.

Anderson, L. A., Wallace, M. N., and Palmer, A. R. (2007). Identification of subdivisions in the medial geniculate body of the guinea pig. *Hearing Research*, 228(1-2):156–167.

Antunes, F. M., Nelken, I., Covey, E., and Malmierca, M. S. (2010). Stimulus-specific adaptation in the auditory thalamus of the anesthetized rat. *PloS One*, 5(11):e14071.

Arnesen, A. R. and Osen, K. K. (1978). The cochlear nerve in the cat: topography, cochleotopy, and fiber spectrum. *Journal of Comparative Neurology*, 178(4):661–678.

Asari, H. and Zador, A. M. (2009). Long-lasting context dependence constrains neural encoding models in rodent auditory cortex. *Journal of Neurophysiology*, 102(5):2638–2656.

Atencio, C. A., Sharpee, T. O., and Schreiner, C. E. (2008). Cooperative nonlinearities in auditory cortical neurons. *Neuron*, 58(6):956–966.

Atencio, C. A., Sharpee, T. O., and Schreiner, C. E. (2009). Hierarchical computation in the canonical auditory cortical circuit. *Proceedings of the National Academy of Sciences*, 106(51):21894–21899.

Bar-Yosef, O. and Nelken, I. (2007). The effects of background noise on the neural responses to natural sounds in cat primary auditory cortex. *Frontiers in Computational Neuroscience*, 1:3.

Bar-Yosef, O., Rotman, Y., and Nelken, I. (2002). Responses of neurons in cat primary auditory cortex to bird chirps: effects of temporal and spectral context. *Journal of Neuroscience*, 22(19):8619–8632.

Bartlett, E. L. and Wang, X. (2005). Long-lasting modulation by stimulus context in primate auditory cortex. *Journal of Neurophysiology*, 94(1):83–104.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London.

Bizley, J. K., Nodal, F. R., Parsons, C. H., and King, A. J. (2007). Role of auditory cortex in sound localization in the midsagittal plane. *Journal of Neurophysiology*, 98(3):1763–1774.

Bizley, J. K. and Walker, K. M. M. (2009). Distributed sensitivity to conspecific vocalizations and implications for the auditory dual stream hypothesis. *Journal of Neuroscience*, 29(10):3011–3013.

Bizley, J. K., Walker, K. M. M., Silverman, B. W., King, A. J., and Schnupp, J. W. H. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *Journal of Neuroscience*, 29(7):2064–2075.

Blake, D. T. and Merzenich, M. M. (2002). Changes of AI receptive fields with sound density. *Journal of Neurophysiology*, 88(6):3409–3420.

Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, 26(3):695–702.

Brosch, M. and Schreiner, C. E. (1997). Time course of forward masking tuning curves in cat primary auditory cortex. *Journal of Neurophysiology*, 77(2):923–943.

Brosch, M. and Schreiner, C. E. (2000). Sequence sensitivity of neurons in cat primary auditory cortex. *Cerebral Cortex*, 10(12):1155–1167.

Brosch, M., Schulz, A., and Scheich, H. (1999). Processing of sound sequences in macaque auditory cortex: response enhancement. *Journal of Neurophysiology*, 82(3):1542–1559.

Buchwald, J., Dickerson, L., and Harrison, J. (1988). Medial geniculate body unit responses to cat cries. In Syka, J. and Masterton, R. B., editors, *Auditory pathway - structure and function*. Plenum Press, New York.

Calabrese, A., Schumacher, J. W., Schneider, D. M., Paninski, L., and Woolley, S. M. N. (2011). A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. *PloS One*, 6(1):e16104.

Calford, M. B. (1983). The parcellation of the medial geniculate body of the cat defined by the auditory response properties of single units. *Journal of Neuroscience*, 3(11):2350–2364.

Calford, M. B. and Aitkin, L. M. (1983). Ascending projections to the medial geniculate body of the cat: evidence for multiple, parallel auditory pathways through thalamus. *Journal of Neuroscience*, 3(11):2365–2380.

Calford, M. B. and Semple, M. N. (1995). Monaural inhibition in cat auditory cortex. *Journal of Neurophysiology*, 73(5):1876–1891.

Cant, N. B. and Benson, C. G. (2003). Parallel auditory pathways: projection patterns of the different neuronal populations in the dorsal and ventral cochlear nuclei. *Brain Research Bulletin*, 60(5-6):457–474.

Casseday, J. H., Fremouw, T., and Covey, E. (2002). The inferior colliculus: a hub for the central auditory system. In Oertel, D., Fay, R. R., and Popper, A. N., editors, *Integrative functions in the mammalian auditory pathway*, pages 238–318. New York: Springer-Verlag.

Chichilinsky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12:199–213.

Christianson, G. B., Sahani, M., and Linden, J. F. (2008). The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *Journal of Neuroscience*, 28(2):446–455.

Clopton, B. M., Winfield, J. A., and Flammino, F. J. (1974). Tonotopic organization: review and analysis. *Brain Research*, 76(1):1–20.

David, S. V., Mesgarani, N., Fritz, J. B., and Shamma, S. A. (2009). Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. *Journal of Neuroscience*, 29(11):3374–3386.

David, S. V., Mesgarani, N., and Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network: Computation in Neural Systems*, 18(3):191–212.

de Boer, E. and de Jongh, H. R. (1978). On cochlear encoding: potentialities and limitations of the reverse-correlation technique. *Journal of the Acoustical Society of America*, 63(1):115–135.

de La Mothe, L. A., Blumell, S., Kajikawa, Y., and Hackett, T. A. (2006). Cortical connections of the auditory cortex in marmoset monkeys: core and medial belt regions. *Journal of Comparative Neurology*, 496(1):27–71.

de Ribaupierre, F. (1997). Acoustical information processing in the auditory thalamus and cerebral cortex. In Ehret, G. and Romand, R., editors, *The central auditory system*, pages 317–398. Oxford University Press, Oxford.

de Ruyter van Steveninck, R. and Bialek, W. (1988). Real-Time Performance of a Movement-Sensitive Neuron in the Blowfly Visual System: Coding and Information Transfer in Short Spike Sequences. *Proceedings of the Royal Society B*, 234(1277):379–414.

Dean, I., Harper, N. S., and McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience*, 8(12):1684–1689.

deCharms, R. C., Blake, D. T., and Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, 280(5368):1439–1443.

Dempster, A., Laird, N., and Rubin, R. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38.

Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, 85(3):1220–1234.

Eggermont, J. J. (1993). Differential effects of age on click-rate and amplitude modulation-frequency coding in primary auditory cortex of the cat. *Hearing Research*, 65(1-2):175–192.

Eggermont, J. J., Johannesma, P. M., and Aertsen, A. M. (1983). Reverse-correlation methods in auditory research. *Quarterly Reviews of Biophysics*, 16(3):341–414.

Elhilali, M., Fritz, J. B., Chi, T.-S., and Shamma, S. A. (2007). Auditory cortical receptive fields: stable entities with plastic abilities. *Journal of Neuroscience*, 27(39):10372–10382.

Englitz, B., Ahrens, M., Tolnai, S., Rübsamen, R., Sahani, M., and Jost, J. (2010). Multilinear models of single cell responses in the medial nucleus of the trapezoid body. *Network: Computation in Neural Systems*, 21(1-2):91–124.

Escabi, M. A. and Schreiner, C. E. (2002). Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *Journal of Neuroscience*, 22(10):4114–4131.

Escola, S., Fontanini, A., Katz, D., and Paninski, L. (2011). Hidden markov models for the stimulus-response relationships of multistate neural systems. *Neural Computation*, 23(5):1071–1132.

Fitzpatrick, D. C., Kuwada, S., Kim, D. O., Parham, K., and Batra, R. (1999). Responses of neurons to click-pairs as simulated echoes: auditory nerve to auditory cortex. *Journal of the Acoustical Society of America*, 106(6):3460–3472.

Flanagan, J. L. (1972). *Speech analysis, synthesis and perception*. Springer-Verlag, New York.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–374.

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, 6(11):1216–1223.

Fritz, J. B., Elhilali, M., and Shamma, S. A. (2007). Adaptive changes in cortical receptive fields induced by attention to complex sounds. *Journal of Neurophysiology*, 98(4):2337–2346.

Gill, P. R., Zhang, J., Woolley, S. M. N., Fremouw, T., and Theunissen, F. E. (2006). Sound representation methods for spectro-temporal receptive field estimation. *Journal of Computational Neuroscience*, 21(1):5–20.

Greene, G., Barrett, D. G. T., Sen, K., and Houghton, C. (2009). Sparse coding of birdsong and receptive field structure in songbirds. *Network: Computation in Neural Systems*, 20(3):162–177.

Grothe, B., Pecka, M., and McAlpine, D. (2010). Mechanisms of sound localization in mammals. *Physiological Reviews*, 90(3):983–1012.

Hackett, T. A., Stepniewska, I., and Kaas, J. H. (1998). Thalamocortical connections of the parabelt auditory cortex in macaque monkeys. *Journal of Comparative Neurology*, 400(2):271–286.

Harris, K. D., Csicsvari, J., Hirase, H., Dragoi, G., and Buzsáki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature*, 424(6948):552–556.

Heil, P. (1997a). Auditory cortical onset responses revisited. I. First-spike timing. *Journal of Neurophysiology*, 77(5):2616–2641.

Heil, P. (1997b). Auditory cortical onset responses revisited. II. Response strength. *Journal of Neurophysiology*, 77(5):2642–2660.

Hunter, I. W. and Korenberg, M. J. (1986). The identification of nonlinear biological systems: Wiener and Hammerstein cascade models. *Biological Cybernetics*, 55(2-3):135–144.

Irvine, D. R. F. (1992). Physiology of the auditory brainstem. In Popper, A. N. and Fay, R. R., editors, *The mammalian auditory pathway: neurophysiology*, pages 153–231. Springer-Verlag, New York.

Jesteadt, W., Bacon, S. P., and Lehman, J. R. (1982). Forward masking as a function of frequency, masker level, and signal delay. *Journal of the Acoustical Society of America*, 71(4):950–962.

Jones, E. G. (1985). *The thalamus*. Plenum Press, New York.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Joris, P. X., Smith, P. H., and Yin, T. C. (1998). Coincidence detection in the auditory system: 50 years after Jeffress. *Neuron*, 21(6):1235–1238.

Kimura, A., Donishi, T., Sakoda, T., Hazama, M., and Tamai, Y. (2003). Auditory thalamic nuclei projections to the temporal cortex in the rat. *Neuroscience*, 117(4):1003–1016.

Klein, D. J., Depireux, D. A., Simon, J. Z., and Shamma, S. A. (2000). Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. *Journal of Computational Neuroscience*, 9(1):85–111.

Klein, D. J., Simon, J. Z., Depireux, D. A., and Shamma, S. A. (2006). Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex. *Journal of Computational Neuroscience*, 20(2):111–136.

Kowalski, N., Depireux, D. A., and Shamma, S. A. (1996a). Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *Journal of Neurophysiology*, 76(5):3503–3523.

Kowalski, N., Depireux, D. A., and Shamma, S. A. (1996b). Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. *Journal of Neurophysiology*, 76(5):3524–3534.

Kuwada, S., Yin, T. C., and Wickesberg, R. E. (1979). Response of cat inferior colliculus neurons to binaural beat stimuli: possible mechanisms for sound localization. *Science*, 206(4418):586–588.

Langner, G. and Schreiner, C. E. (1988). Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *Journal of Neurophysiology*, 60(6):1799–1822.

Lee, C. C., Imaizumi, K., Schreiner, C. E., and Winer, J. A. (2004). Concurrent tonotopic processing streams in auditory cortex. *Cerebral Cortex*, 14(4):441–451.

Lee, C. C. and Sherman, S. M. (2010). Drivers and modulators in the central auditory pathways. *Frontiers in Neuroscience*, 4(1):79–86.

Lewicki, M. S. (1998). A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):53–78.

Linden, J. F., Liu, R. C., Sahani, M., Schreiner, C. E., and Merzenich, M. M. (2003). Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. *Journal of Neurophysiology*, 90(4):2660–2675.

Lomber, S. G. and Malhotra, S. (2008). Double dissociation of 'what' and 'where' processing in auditory cortex. *Nature Neuroscience*, 11(5):609–616.

Lomber, S. G., Meredith, M. A., and Kral, A. (2010). Cross-modal plasticity in specific auditory cortices underlies visual compensations in the deaf. *Nature Neuroscience*, 13(11):1421–1427.

Lomber, S. G., Payne, B. R., and Horel, J. A. (1999). The cryoloop: an adaptable reversible cooling deactivation method for behavioral or electrophysiological assessment of neural function. *Journal of Neuroscience Methods*, 86(2):179–194.

Machens, C. K., Wehr, M. S., and Zador, A. M. (2004). Linearity of cortical receptive fields measured with natural sounds. *Journal of Neuroscience*, 24(5):1089–1100.

Malmierca, M. S., Cristaudo, S., Pérez-González, D., and Covey, E. (2009). Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *Journal of Neuroscience*, 29(17):5483–5493.

Margoliash, D. and Fortune, E. S. (1992). Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. *Journal of Neuroscience*, 12(11):4309–4326.

Marmarelis, P. Z. and Marmarelis, V. Z. (1978). *Analysis of physiological systems: the white-noise approach*. New York: Plenum.

May, B. J. (2000). Role of the dorsal cochlear nucleus in the sound localization behavior of cats. *Hearing Research*, 148(1-2):74–87.

McAlpine, D. and Grothe, B. (2003). Sound localization and delay lines–do mammals fit the model? *Trends in Neurosciences*, 26(7):347–350.

Miller, L. M., Escabi, M. A., Read, H. L., and Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, 87(1):516–527.

Moore, B. C. (1980). Mechanism and frequency distribution of two-tone suppression in forward masking. *Journal of the Acoustical Society of America*, 68(3):814–824.

Morest, D. K. (1965). The laminar structure of the medial geniculate body of the cat. *Journal of Anatomy*, 99:143–160.

Narendra, K. S. and Gallman, P. G. (1966). An iterative method for the identification of nonlinear systems using a Hammerstein model. *IEEE Transactions on Automatic Control*, 11(3):546–550.

Nelken, I., Kim, P. J., and Young, E. D. (1997). Linear and nonlinear spectral integration in type IV neurons of the dorsal cochlear nucleus. II. Predicting responses with the use of nonlinear models. *Journal of Neurophysiology*, 78(2):800–811.

Noreña, A. J., Gourévitch, B., Pienkowski, M., Shaw, G., and Eggermont, J. J. (2008). Increasing spectrotemporal sound density reveals an octave-based organization in cat primary auditory cortex. *Journal of Neuroscience*, 28(36):8885–8896.

Oertel, D. (1999). The role of timing in the brain stem auditory nuclei of vertebrates. *Annual Review of Physiology*, 61:497–519.

Okatan, M., Wilson, M. A., and Brown, E. N. (2005). Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Computation*, 17(9):1927–1961.

Paninski, L. (2003). Convergence properties of three spike-triggered analysis techniques. *Network: Computation in Neural Systems*.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262.

Paninski, L., Pillow, J. W., and Simoncelli, E. P. (2004). Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. *Neural Computation*, 16(12):2533–2561.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilinsky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999.

Polley, D. B., Read, H. L., Storace, D. A., and Merzenich, M. M. (2007). Multiparametric auditory receptive field organization across five cortical fields in the albino rat. *Journal of Neurophysiology*, 97(5):3621–3638.

Rhode, W. S., Oertel, D., and Smith, P. H. (1983). Physiological response properties of cells labeled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus. *Journal of Comparative Neurology*, 213(4):448–463.

Rouiller, E., de Ribaupierre, Y., Toros-Morel, A., and de Ribaupierre, F. (1981). Neural coding of repetitive clicks in the medial geniculate body of cat. *Hearing Research*, 5(1):81–100.

Rust, N. C., Schwartz, O., Movshon, J. A., and Simoncelli, E. P. (2004). Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey V1. *Neurocomputing*, 58:793–799.

Rutkowski, R. G., Shackleton, T. M., Schnupp, J. W. H., Wallace, M. N., and Palmer, A. R. (2002). Spectrotemporal receptive field properties of single units in the primary, dorsocaudal and ventrorostral auditory cortex of the guinea pig. *Audiology & Neurotology*, 7(4):214–227.

Sadagopan, S. and Wang, X. (2009). Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. *Journal of Neuroscience*, 29(36):11192–11202.

Sahani, M. (1999). *Latent variable models for neural data analysis*. PhD thesis, California Institue of Technology.

Sahani, M. and Linden, J. F. (2003a). Evidence optimization techniques for estimating stimulus-response functions. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in neural information processing systems*, pages 301–308. MIT Press, Cambridge, MA.

Sahani, M. and Linden, J. F. (2003b). How linear are auditory cortical responses? In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in neural information processing systems*, pages 109–116. MIT Press, Cambridge, MA.

Schreiner, C. E. and Langner, G. (1988). Periodicity coding in the inferior colliculus of the cat. II. Topographical organization. *Journal of Neurophysiology*, 60(6):1823–1840.

Schwartz, O., Pillow, J. W., Rust, N. C., and Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of Vision*, 6(4):484–507.

Sen, K., Theunissen, F. E., and Doupe, A. J. (2001). Feature analysis of natural sounds in the songbird auditory forebrain. *Journal of Neurophysiology*, 86(3):1445–1458.

Sharpee, T. O., Rust, N. C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Computation*, 16(2):223–250.

Shoham, S., Paninski, L. M., Fellows, M. R., Hatsopoulos, N. G., Donoghue, J. P., and Normann, R. A. (2005). Statistical encoding model for a primary motor cortical brain-machine interface. *IEEE Transactions on Biomedical Engineering*, 52(7):1312–1322.

Smith, A. L., Parsons, C. H., Lanyon, R. G., Bizley, J. K., Akerman, C. J., Baker, G. E., Dempster, A. C., Thompson, I. D., and King, A. J. (2004). An investigation of the role of auditory cortex in sound localization using muscimol-releasing Elvax. *European Journal of Neuroscience*, 19(11):3059–3072.

Smith, P. H. and Populin, L. C. (2001). Fundamental differences between the thalamocortical recipient layers of the cat auditory and visual cortices. *Journal of Comparative Neurology*, 436(4):508–519.

Steinberg, L. J. and Peña, J. L. (2011). Difference in response reliability predicted by spectrotemporal tuning in the cochlear nuclei of barn owls. *Journal of Neuroscience*, 31(9):3234–3242.

Stiebler, I., Neulist, R., Fichtel, I., and Ehret, G. (1997). The auditory cortex of the house mouse: left-right differences, tonotopic organization and quantitative analysis of frequency representation. *Journal of Comparative Physiology A*, 181(6):559–571.

Tan, A. Y. Y., Zhang, L. I., Merzenich, M. M., and Schreiner, C. E. (2004). Tone-evoked excitatory and inhibitory synaptic conductances of primary auditory cortex neurons. *Journal of Neurophysiology*, 92(1):630–643.

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A. S., Vinje, W. E., and Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12(3):289–316.

Theunissen, F. E., Sen, K., and Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, 20(6):2315–2331.

Tollin, D. J. (2003). The lateral superior olive: a functional role in sound source localization. *The Neuroscientist*, 9(2):127–143.

Ulanovsky, N., Las, L., Farkas, D., and Nelken, I. (2004). Multiple time scales of adaptation in auditory cortex neurons. *Journal of Neuroscience*, 24(46):10440–10453.

Ulanovsky, N., Las, L., and Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, 6(4):391–398.

Valentine, P. A. and Eggermont, J. J. (2004). Stimulus dependence of spectro-temporal receptive fields in cat primary auditory cortex. *Hearing Research*, 196(1-2):119–133.

Volterra, V. (1930). *Theory of functionals and of integral and integro-differential equations*. Glasgow, UK: Blackie and Son.

Von Békésy, G. (1980). *Experiments in hearing*. New York: McGraw-Hill.

Wang, X., Lu, T., Snider, R. K., and Liang, L. (2005). Sustained firing in auditory cortex evoked by preferred stimuli. *Nature*, 435(7040):341–346.

Wehr, M. and Zador, A. M. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature*, 426(6965):442–446.

Wehr, M. and Zador, A. M. (2005). Synaptic mechanisms of forward suppression in rat auditory cortex. *Neuron*, 47(3):437–445.

Wiener, N. (1958). *Nonlinear problems in random theory*. New York: Wiley.

Winer, J. A., Miller, L. M., Lee, C. C., and Schreiner, C. E. (2005). Auditory thalamocortical transformation: structure and function. *Trends in Neurosciences*, 28(5):255–263.

Woolley, S. M. N., Fremouw, T. E., Hsu, A., and Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, 8(10):1371–1379.

Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Reviews of Neuroscience*, 29:477–505.

Young, E. D., Spirou, G. A., Rice, J. J., and Voigt, H. F. (1992). Neural organization and responses to complex stimuli in the dorsal cochlear nucleus. *Philosophical Transactions of the Royal Society of London B*, 336(1278):407–413.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579.