Modelling the neural coding of natural sounds in the auditory cortex

Monzilur Rahman

Keble College, University of Oxford

Thesis submitted for the degree of

Doctor of Philosophy

Hilary 2021

To my parents –

Mariam Begum and Jahur Uddin,

Who were my first teachers during those home-schooling years

Who ingrained the value of scholarship within me very early on

And continue to encourage my academic endeavors to this day

And to my siblings -

Mastofa Hosen, who stood by my side against all odds

Mamunur Rashid, who had higher hopes about me than I had for myself

Joyita Noasheen, who thought I could make anything possible.

It is their belief in me that keeps me going forward.

Quotes

“I can live with doubt, uncertainty and not knowing. … I don’t feel frightened by not knowing things.” - Richard P. Feynman

“Any man could, if he were so inclined, be the sculptor of his own brain.” - Santiago Ramon y Cajal

"All models are wrong, but some are useful" - George E. P. Box

Abstract

When neurons fire in the auditory cortex what do they represent? What is the form of transformation that sound stimuli go through in order to produce activity in the auditory cortex? Studying how neural activity changes in response to a stimulus parameter, e.g. the level or frequency of sound, has been a way to address such questions in the past. In more recent times, an alternative approach has been to measure the responses of cortical neurons to natural stimuli. Here, we investigate the responses of neurons in ferret auditory cortex, recorded using extracellular electrode probes, to a diverse collection of natural sounds using models that aim to capture many aspects of the information transformation process from the ear to the cortex. We use the method of encoding - whereby we build functional transformations that predict neural responses from the sound stimulus, and also of decoding - whereby we build functional transformations that estimate the stimulus from neural responses. Our primary goal has been to focus our modelling effort to produce mechanistic and interpretive insights into auditory processing in the brain. Firstly, we examine to what extent biological details matter in making a model of the auditory periphery that provides input to the cortex. Secondly, we study possible biological mechanisms underlying the temporal receptive fields of auditory cortical neurons. Thirdly, we ask whether neural populations in the auditory cortex represent the past or future of ongoing sound stimuli. We found that simple models are well suited for models of auditory encoding in the brain, and that stimulus history in linear receptive fields of auditory neurons can be explained by a dynamic network model, and also that a population of neurons might contain non-linearities that enable them to predict the future. These results will improve our understanding of some of the fundamental questions relating to neural coding in the auditory system.

Acknowledgement

This DPhil thesis summarizes the experiments and analyses that I did during the past four and a half years. However, the description of my research work would be incomplete without mentioning all the wonderful human beings that have been involved in various ways in guiding and supporting me in this process. In particular, it was not easy to keep going during a pandemic. But I am blessed to be helped and supported by so many kind and generous souls.

First and foremost, I would like to acknowledge my supervisors Nicol Harper, Ben Willmore and Andy King. You have been true guides and mentors. My research has been immensely impacted by your invaluable support and feedback that you have provided throughout. Your advice has helped me develop as a matured person and a seasoned researcher and for that I am grateful to you. Thank you for the immense amount of work that you have put in reviewing my works which has shaped this thesis. Andy, thank you for providing me space in your lab and always making sure that I had the resources I needed for my research. Ben, thank you for providing me advice on research and outside-of-research topics. Nicol, special thanks to you for you have introduced me to the world of encoding models in auditory neuroscience and you have helped in conceiving the dynamic network model. Your guidance has aided me in developing my skills as a theoretician. Thank you all for being inspirational role models.

I would not have been able to survive the desolating times, especially, during the pandemic, without the support of Aleksandar Ivanov and Tai-Ying Lee. Our tea club, book club, journal club and our attending conferences together became a very valuable part of my DPhil. You have been the most supportive colleagues and true friends. In many ways your feedback and guidance shaped my DPhil and the thesis. Alex and Tai, having tea with you in the park, planning the Japanese tea ceremony, the hot pot will be some of the most memorable things of my DPhil period. Tai, thanks for bringing me to all the life drawing sessions, music fests and various other

cheerful things. Alex, thanks always for your advice and support. I am grateful to both of you and to have you as my friends.

I would also like to thank all other members of ANG. Thanks to Freddy Trinh and Kat Poole for those walks and chats in the park in the early days of my DPhil and special thanks to Chris Breen for being a source of support and a great friend during the past year.

I would also like to extend my special thanks to trusted colleague and dear friend, Peter Hasenhuetl of CNCB, Oxford for your valuable guidance and advice. I look forward to more collaboration with you in future.

I would like to thank all the senior colleagues and scientists that I have received valuable guidance and feedback from during my DPhil, notably, Kerry Walker and Tim Vogels for their valuable feedbacks during the transfer of status and confirmation of status examinations, Armin Lak for being a trusted guide and mentor outside of my lab, Matteo Carandini, Cristina Savin, Simon Butt, Eero Simoncelli, Anthony Movshon, Simon Laughlin and Jennifer Linden for guiding me and providing me valuable advice that shaped my DPhil training in many ways.

I would like to pay special thanks to Crystal Leung who was involved in part of a collaborative project that later became a chapter of this thesis.

I thank the CAJAL course in computational neuroscience for providing me with a great training program in Lisbon. I have learned so much during that one month of summer break. I would also like to thank all my funders – the Clarendon Fund Scholarship and Keble College. I have also been supported by various organizations; the Keble Association, the Association for Research in Otolaryngology, Action on Hearing Loss and Boehringer Ingelheim Fonds that provided me with travel grants to attend numerous conferences which was an integral part of my DPhil training.

I would like to thank the Department of Physiology, Anatomy and Genetics for providing all the training, resources, and seminars during the course of my DPhil.

I would also like to thank Keble College and specially the MCR for filling my DPhil period with countless non-academic experiences. I really enjoyed being a part of the MCR committee as well.

My time in Oxford was highly enriched by Mohd Anisul Karim, Amita Farzana, Hassand Saad Ifti and Shamir Montazid. Your company has added color to the past several years. Our regular hangouts, parties, and café visits provided the positivity that I needed to keep going. You kept me recharged during the time of burn outs and gave me the support I needed. I will never forget our Amsterdam tour. Our CoronaCast became a social media sensation during the CoVID pandemic. MAK and Amita got married recently - I look forward to your wedding celebration in summer when the pandemic is over, and things become normal.

I am grateful to Syed Muntasir Mamun for being a trusted friend and longstanding guide. You have been by my side in some of the most difficult periods for which I will always be grateful to you. Your wisdom and intellect guided me always. Your energy inspired me. I thank you for being my friend.

I would like to thank Kristina Astrom for being a part of my life. I enjoyed travelling, partying, and spending time with you for the last couple of years. Specially, the Spain tour, going to balls and our Stockholm visit - these surely have been some of the most memorable moments of my DPhil period. I look forward to more joy and happiness with you in future and more adventures when things get back to normal.

All other people whose path I crossed or who crossed my path during my DPhil and whose names are not mentioned here - many of them, although for a short time, have influenced my life immensely, and shifted my ideas about life - I thank you all.

Finally, I am vastly indebted to my family who supported me throughout.

Contribution statement

Part of Chapter 4 of this thesis was initiated as an MSc project (Hilary Term 2018, MSc in Neuroscience, University of Oxford) of Crystal Y. C. Leung jointly supervised by Monzilur Rahman and Nicol S. Harper. Some of the preliminary results were obtained during that project from a dataset of 73 auditory cortical neurons (A1) in anesthetized ferrets.

Table of Contents

Chapter 1 | Introduction

In neuroscience, the concept of neural coding is often used to denote what information an individual neuron or a population of neurons encodes and how that information is encoded (see (Johnson, 2000) for a brief review on neural coding). Since neurons are the fundamental units of brain function, many questions in neuroscience often begin with the study of neural coding. For example, the study of sensory systems often asks how a particular stimulus generates specific responses in the brain and the study of motor systems asks how specific motor sequences are triggered by particular neural activity patterns.

While it is challenging to probe the nature of information encoded by higher cognitive areas simply because the appropriate input may be unknown, sensory systems provide an opportunity to probe the system due to the more direct relationship of the neural responses to external stimuli. As a result, sensory systems are well suited to study neural coding in a detail that higher brain areas are not. Many early discoveries in neuroscience have resulted from wondering about what real-world stimuli are represented in the brain and how. In fact, many of these studies have investigated the responses of sensory areas to simple stimuli.

In case of the visual system, for example, valuable insights have been obtained by studying the responses of neurons at various levels of the visual hierarchy in response to different simple synthetic stimulus. Some seminal studies include studying the responses of retinal ganglion cells in response to spots of light (Hartline, 1940), studying the responses of neurons in thalamus and cortex in response to oriented bars (Hubel and Wiesel, 1962). These studies often aimed at finding the stimuli that best stimulate a particular brain area, and resulted in important insights about the

underlying cellular components of information encoding – for example, the discovery of simple and complex cells through the investigation of the responses to oriented bars and drifting gratings.

A more recent avenue of research has been to study the responses of neurons to natural stimuli (Passaglia et al., 1997; Gallant et al., 1998). Work on the visual system has inspired similar approaches to the auditory system. Studying the activity of auditory neurons in response to sound stimuli, for example, pure tones (Woolley and Casseday, 2004) and natural sounds (Margoliash and Konishi, 1985; Nelken et al., 1999; Theunissen et al., 2001) has produced valuable insights in the understanding of this system.

Although significant advances have been made in the understanding of neural coding in the brain and models have been developed to predict the time course of neural responses, we are still far from finding the perfect models. We are still in search of models that can describe the responses of neurons and relate them to the underlying architecture and dynamics of the network of neurons. Advances in neural recording techniques have made it possible to record simultaneously from numerous neurons, from anesthetized, awake or behaving animals and in response to sensory stimulation. Investigating these dense recordings regarding neural coding requires a combination of theory and advanced computational methods. In this thesis, we will study the responses of neurons in the auditory pathway to naturalistic stimuli by employing newly developed statistical methods and models with the aim of improving our understanding of neural coding in the auditory system. But first, we will provide a brief introduction to the relevant concepts.

1.1 Modelling the responses of neurons

The complexity and scale of biological networks of neurons make them challenging to understand, firstly because it is difficult to collect data from a sufficient number of neurons, and secondly because it is difficult to analyze the data collected from a small sub-sample of the network and, from this, infer the function of the whole network. Modelling solves this problem to some extent by searching for models that are generalizable across datasets and across sample of neurons. Modelling frameworks thus allow for the investigation of information processing principles and biological mechanisms present in biological neural networks.

While one objective of building a model of a system can be to understand in detail the underlying mechanisms that constitute that system, models can also abstract away some features of the underlying mechanism and describe the system with the level of detail that is necessary for the question of interest. In his seminal work, David Marr (1982) provided a solution for this dilemma of whether to study the abstract aspect or to study the mechanistic detail of a system. He advocated for three levels of understanding for the brain or any other information processing device: the computations performed, the algorithms used, and the hardware implementation of those algorithms (Marr, 1982). Determining which details are particularly relevant for the function of a neural circuit is crucial to understanding it. On one end of the spectrum is the computational level, understanding the role performed by the circuit. In the middle are algorithm level that essentially involve describing the computational transformation that distills from the complex neural circuit. On the other end of the spectrum is the implementation level, which address the details of the neural circuitry by which the algorithm is implemented. Being able to define the appropriate level of modelling can be very useful while modelling biological systems, in this case, neurons responding to stimuli.

A model can often fall into one of three categories that often serve three related but separate purposes (Dayan and Abbott, 2001). These three categories can be loosely related to Marr's level of analysis. We will put these purposes in the context of the study of neural responses:

1. Descriptive: Understanding the response properties of an individual neuron and a network of neurons. This can be related to Marr's algorithm level.
2. Mechanistic: Identifying the mechanisms underlying the neuron's stimulus-response function. This can be related to Marr's implementation level.
3. Interpretive: Understanding of functions of neurons and neural systems for the animal, and understanding why the system behaves as it does. This can be related to Marr's computational level.

In this thesis, we will mostly focus on descriptive models, although they are inspired by mechanistic concepts (particularly in Chapters 2 and 3) and are informative about interpretations (particularly in Chapter 4). Our effort towards modelling the responses of neurons to natural sounds can be seen to be aimed at providing a descriptive understanding of the auditory system. However, by investigating the role of particular biological mechanisms in this description, we can widen the scope of the models so that they become more mechanistic. Likewise, we can ask various 'why' questions by widening the scope of the descriptive models to make them more interpretive in nature.

1.2 The auditory system

Vibrating objects can transmit energy through pressure waves in a medium, such as a gas or liquid. The auditory system receives this energy which we call sound and converts it into a percept

that we call hearing. The auditory system, like other sensory systems, extracts useful information about the environment from this sensory input. This system is important for detecting, identifying and localizing objects and events (Bregman, 1990; Alain et al., 2001; Griffiths and Warren, 2004; Nelken et al., 2014), especially ones that are outside of the field of vision, and for communication (Bennur et al., 2013; Prather, 2013). Humans in particular rely on the auditory system for communication through speech (Hauser et al., 2002; Greenberg and Ainsworth, 2006). Many other animals also communicate through their vocalizations (Thorpe, 1972). The information processing that is required to carry out these very sophisticated functions takes place in various structures and neural circuits in the brainstem and cortex. Before delving into the detail of modelling the responses of the auditory system, we will go through a brief description of the neuronal pathway that makes up this system.

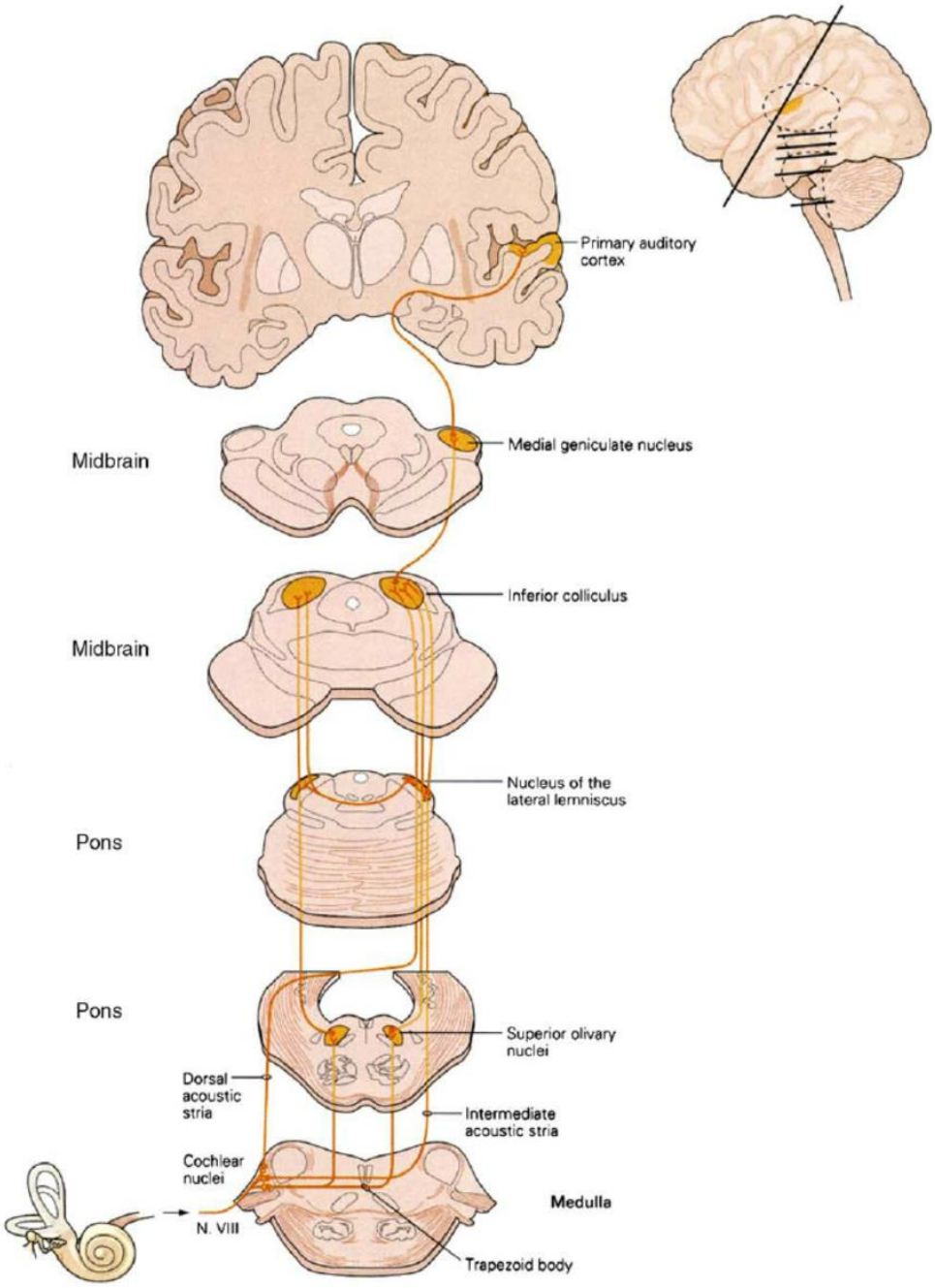

Figure 1.1: The auditory pathway originates at the ear and extends up to the cortical regions. *The ear contains an organ called the cochlea that transduces auditory signal into the action potentials of the cochlear nerves. The signal coming from the cochlear nerve is successively transmitted through multiple nuclei to the cortex. The first of the nuclei is the cochlear nucleus located in the medulla, which receives the output of the cochlear nerves. The cochlear*

nucleus sends its output to the superior olivary nucleus in the pons, which in turn sends its output to the nuclei of lateral lemniscus and the inferior colliculus, which is situated in the midbrain. The inferior colliculus sends its output to the medial geniculate body of the thalamus which projects to the cortex (Figure adapted from (Hudspeth et al., 2013)).

1.2.1 Auditory periphery

The auditory periphery provides the interface between the auditory system and the external world and is composed of the ear and the auditory nerve. Classically, the ear is described as consisting of three separate sections - outer ear, middle ear and inner ear. The principal structures of the outer ear are the pinna and concha and the ear canal, which ends at the eardrum. The eardrum forms the boundary between the outer ear and the middle ear. The middle ear is an air-filled structure with three small bones or ossicles – the malleus, incus and stapes. The last of these bones, the stapes, makes physical contact with a membrane known as the oval window, behind which lies the cochlea of the inner ear. The cochlea is a spiral-shaped organ that along with the vestibular periphery forms the inner ear. We are not going to discuss the vestibular system here. The cochlea is divided longitudinally into three fluid-filled compartments, known as scalae, which are separated by two membranes. The organ of Corti is situated on one of these membranes – the basilar membrane – and this comprises the inner and outer hair cells (and other supporting cells) that work as the transducers of mechanical energy into electrical energy (Schnupp et al., 2011).

Collectively, the structures in the outer, middle and inner ear are involved in detecting, differentiating and transducing sound signals into electrical signals, and transmitting them to the

auditory nerve (Robles and Ruggero, 2001). The outer ear funnels the fluctuations of air pressure that compose the sound to the eardrum or tympanic membrane, which vibrates the bony ossicles of the middle ear. Theses vibrations are in turn coupled via the oval window to the inner ear. This results in displacement of the basilar membrane and surrounding fluids. Because different frequency components of sounds cause maximum displacement at different points along the length of the basilar membrane, this process allows the basilar membrane to decompose sounds into frequency components, with high frequencies preferentially vibrating the base of the membrane and low frequencies at the apex (Rhode, 1971; Evans and Wilson, 1975; Ehret, 1978; Ruggero and Rich, 1986; Ruggero, 1992). The inner hair cells transduce the vibration of the basilar membrane into fluctuations in membrane potential, which modulate the firing of action potentials in the auditory nerve (Robles and Ruggero, 2001). These signals are transmitted by the auditory nerve to the cochlear nucleus. Due to the mechanical tuning of the basilar membrane, different auditory nerve fibers have different characteristic frequencies (Narayan et al., 1998). There is a non-linear compressive relationship between sound level and the firing rate of auditory nerve fibers (Sachs and Abbas, 1978; Schlauch et al., 1998). This may be a factor in the approximately logarithmic relationship between sound intensity and its perceived loudness (Colburn et al., 2003).

Each auditory nerve fiber is tuned for sound frequency since it receives synaptic input from a hair cell at a particular position on the basilar membrane that responds maximally to that particular sound frequency (Rhode, 1971; Evans and Wilson, 1975; Ehret, 1978; Ruggero and Rich, 1986; Ruggero, 1992). This process has some characteristics in common with a Fourier decomposition of sound signals (Nilsson and Møller, 1977; von Békésy and Peake, 1990). This means that each auditory nerve fiber has a preferred frequency, to which it responds more strongly relative to other frequencies. Another way to describe the responsiveness of a nerve fiber is by using its impulse

response (Recio et al., 1998; Recio and Rhode, 2000) – the response to a click. The impulse response function can provide valuable insight about the temporal aspect of the response (Recio et al., 1998; Recio and Rhode, 2000). Furthermore, auditory nerve fibers with similar frequency preferences can have different spiking rates based on their thresholds – the minimum sound level to which they respond. This is the basis for classifying the fibers into low spontaneous rate, medium spontaneous rate and high spontaneous rate fibers (Winter et al., 1990; Yates et al., 1990). The spontaneous rate of a fiber is inversely related to its threshold – the higher the threshold the lower the spontaneous rate.

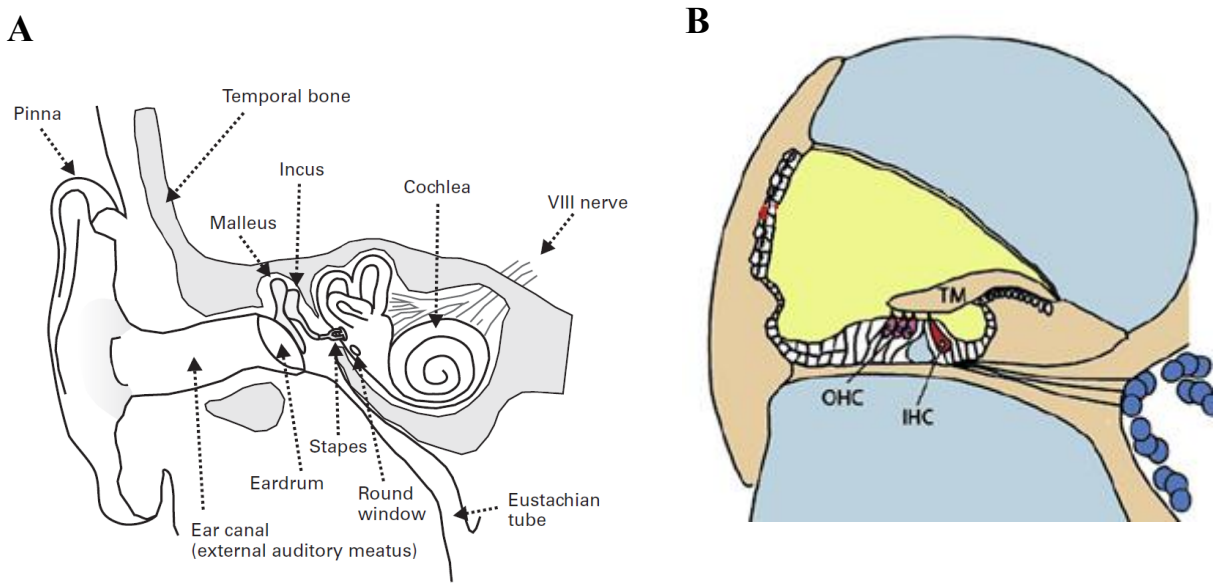

Figure 1.2: Structure of the ear and cochlea. *A. A cross section of human ear showing different parts of the ear (Figure adapted from (Schnupp et al., 2011)). B. A cross section of the feline cochlea. IHC – inner hair cells, OHC – outer hair cells, TM, tectorial membrane [Figure adapted from (Ryugo and Menotti-Raymond, 2012)).*

1.2.2 The brainstem and auditory pathway to cortex

In anatomical detail, the auditory pathway from the auditory nerve to the cortex is composed of various nuclei. These nuclei are the locations of synapses between neurons and consist of numerous cell types. Along with ascending and descending fibers, there are many types of interneurons. The first such nucleus in the auditory pathway is the cochlear nucleus (CN). The CN is a complex structure that can be subdivided into various parts (e.g. anterior, posterior, dorsal and ventral). These regions can be distinguished physiologically and to some extent are constituted with different cell types. Cell types of CN include spherical bushy cells, stellate cells, globular bushy cells, fusiform cells and octopus cells (Osen, 1969; Rubio, 2018; Trussell and Oertel, 2018). These cells have diverse spectral and temporal response properties (Rhode and Smith, 1986; Schnupp et al., 2011). For example, bushy cells carry the precise temporal pattern of firing of auditory nerves while stellate cells do not carry the precise temporal pattern but are highly sensitive to spectral cues. CN also contains cells that are activated by the onset of sound and cells that are inhibited by the onset of sound – which may play role in the encoding of auditory contrast (Schnupp et al., 2011). These cells make extensive inter-connections and some of them send output to the ascending pathway. The outputs from the CN are first relayed to superior olivary nuclei (SOC) and then to a region of the auditory midbrain called the inferior colliculus (IC), or go directly to the IC. Some ascending fibers from the SOC form branches in the nucleus of lateral lemniscus (NLL) on their way to the IC. The IC have various sub-regions and with many different cell types (Aitkin et al., 1975; Peruzzi et al., 2000; Ito et al., 2016; Chen et al., 2018; Schofield and Beebe, 2019). The IC neurons are topographically organized into frequencies that play role in the spectral analysis of sound, binaural processing and level dependent adaptation (Aitkin et al., 1975; Dean et al., 2005, 2008; Escabí and Read, 2005; Schreiner and Winer, 2007; Wen et al., 2009). Projections from the IC are sent to the medial geniculate body (MGB) of the

thalamus. While the ventral part of the MGB mainly relays the topographically organized sound representation from the IC to the cortex, other parts of the MGB seem to be involved in more complex interactions and processing (Vasquez-Lopez et al., 2017; Gilad et al., 2020).

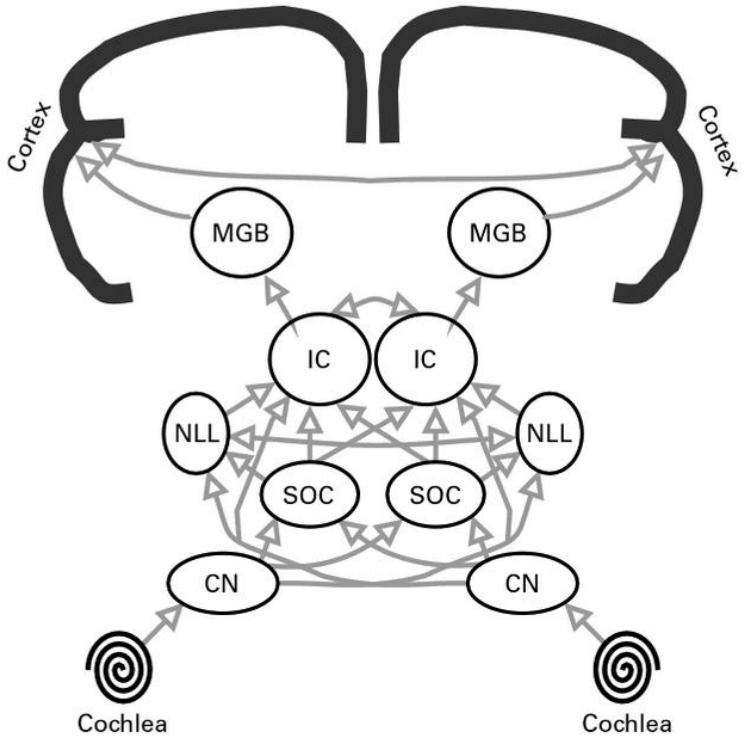

Figure 1.3: A schematic showing the connectivity of the relays of the ascending auditory pathway from the cochlea to the cortex. *This is also called the lemniscal auditory pathway. CN – cochlear nucleus, SOC – superior olivary nucleus, NLL – nucleus of lateral lemniscus, IC – inferior colliculus, MGB – medial geniculate body (Figure adapted from (Schnupp et al., 2011)).*

1.2.3 Auditory cortex

The auditory cortex is the cortical region that is primarily involved in processing the auditory inputs coming from subcortical regions. This area of the brain is divided into various auditory cortical

fields. These brain areas receive inputs from the thalamus as well as from other cortical regions and send outputs to them and to other brain regions (Read et al., 2002; Bizley et al., 2005). A characterizing feature of the auditory cortex is that neurons in some cortical fields are topographically organized according to their sound frequency selectivity (Oonishi and Katushi, 1965; Abeles and Goldstein, 1970; Shamma et al., 1993; Recanzone et al., 1999; Bizley et al., 2005). This topography is called tonotopy, and is present in every stage of the lemniscal auditory pathway (Schreiner and Winer, 2007). Tonotopy is an important basis for dividing the auditory cortex into primary, secondary and higher order cortical areas – primary being highly tonotopic, secondary being somewhat tonotopic and other higher order areas mostly lacking the presence of tonotopic gradients (Schnupp et al., 2011). In fact, tonotopy is often used during experiments to establish the boundary of cortical fields. Other criteria in deciding the boundary of cortical fields include frequency tuning, latency, anatomical organization and connectivity of a cortical field (Bizley et al., 2005). The structural organization of these fields are often species specific (Schnupp et al., 2011). The primary auditory cortex in a hierarchical sense is the first station of these cortical areas along the central auditory pathway, receiving input from the MGB of the thalamus although secondary areas might also receive input directly from the MGB (Winer et al., 2005).

While the coarse structure of the auditory cortex, at least in the primary areas, can be highly organized into tonotopy, at a finer scale the structure could be more heterogeneous (Kanold et al., 2014; Gaucher et al., 2020). Moreover, the tuning properties of individual neurons can be much more complex than just being selective to a single sound frequency (Sutter and Schreiner, 1991; Kadia and Wang, 2003; Sadagopan and Wang, 2009; Gaucher et al., 2020). Furthermore, while the ascending auditory pathway sends input from subcortical areas to the auditory cortex, there are also numerous descending connections. In fact, all sub-cortical areas down to the CN can receive top-down inputs (Schnupp et al., 2011).

The principal type of excitatory neurons in the auditory cortex is the glutamatergic pyramidal neurons. There are also numerous types and subtypes of inhibitory interneurons. These are mainly GABAergic. Some examples are, parvalbumin-positive interneurons, somatostatin positive interneurons and vasopressin positive interneurons (see (Blackwell and Geffen, 2017) for a review). Mammalian cortex is laminar in structure, usually consisting of six layers - the distribution of different cell types and their connectivity is influenced by their position in these cortical layers (Douglas and Martin, 2004). Auditory cortex, like other sensory cortices, receives most of the thalamocortical inputs to neurons in the middle layers. This information primarily travels to the superficial layers and then to the deeper layers, which then send projections out of the auditory cortex. Superficial layers also send extensive connections to lateral areas of the cortex (Mitani et al., 1985; Linden, 2003). There are also numerous other interconnections within and between layers. The location of a neuron in the cortical layers seems to influence the representation of information (Wallace and Palmer, 2008; Christianson et al., 2011).

1.3 The value of studying the responses of auditory neurons to natural stimuli

Although simple well-defined artificial stimuli, such as tones or gratings, can produce valuable insights about a system, it is necessary to probe a system with natural stimuli since it is likely that it is the input regime that defines the functional optima of that system. In the case of the auditory system, the natural input is sounds that are experienced by animals in their natural environment, for example animal calls, footsteps, rushing water, wind in the trees, and so on. So, natural sounds are a collection of sound stimuli that are likely to be experienced with higher probability than other sounds. Moreover, higher brain areas often respond to natural stimuli but not to simple, synthetic

stimuli (Suga et al., 1978; Margoliash and Konishi, 1985; Margoliash and Fortune, 1992). There are also many context-dependent aspects to the neural responses that are better elicited by natural stimuli than by simple synthetic stimuli (Nelken et al., 1999). Furthermore, response properties of neurons in the auditory cortex estimated using natural stimuli seem to be different from when estimated from synthetic stimuli (Laudanski et al., 2012).

1.4 Encoding models and spectro-temporal receptive fields of auditory neurons

Determining the response properties of neurons and relating them to their biological characteristics is a longstanding research avenue. By observing the responses of neurons to simple stimuli, important insights have been gathered about the auditory system by early research (Oonishi and Katushi, 1965; Abeles and Goldstein, 1970). In more recent times, this research avenue has been advanced by applying modern statistical and machine learning techniques to build models of neural responses. As discussed in a previous section, modelling can improve the descriptive and mechanistic understanding of a system. These two purposes are often served by encoding models. Encoding models are models that predict the responses of neurons for a given input. For auditory neurons, encoding models are used to predict the responses of neurons to sound stimuli. While responses to simple stimuli like pure tones can be studied by estimating a tuning curve for a neuron, the study of more complex stimuli, such as white noise or natural sounds, demands methods that can relate numerous parameters of the stimuli to the responses of neurons. The equation $r = E(s)$ describes a very general encoding model. If a stimulus parameter, s results in neural response, r, then the encoding function, E relating these two quantities represents the encoding model of the neural response.

Some of the encoding model studies of the auditory system were carried out in modelling the response properties of cochlea and auditory nerve in response to sound inputs. These early studies used reverse correlation to study the relationship between white noise input and the response of cochlea and auditory nerve (De Boer, 1969; Boer, 1975; de Boer, 1991). The model parameters that were important to study these auditory peripheral structures included frequency selectivity and latency. These modelling studies ultimately led to the development of spectro-temporal receptive field (STRF) models for auditory periphery. In an STRF model, the response property of a neuron or nerve fiber is expressed as the linear relationship between neural response and sound stimuli at frequencies and latencies (Aertsen et al., 1980; Aertsen and Johannesma, 1981a, 1981b) (more detail about STRF models follows). In more recent time, similar models have been applied to study neurons in cochlear nucleus (Bandyopadhyay et al., 2007; Tan et al., 2015). These early studies formed the basis for modelling the auditory periphery and other auditory areas in the brain.

Sensory neurons are often characterized using the so-called concept of receptive fields. For neurons in the auditory cortex, linear spectrotemporal receptive fields (STRFs) have been used as a simple model of their responses. Neurons' responsiveness to specific spectral content of sounds with a specific latency is determined by finding the underlying linear relationship between a neuron's response and its auditory input. STRFs (Aertsen et al., 1980; Aertsen and Johannesma, 1981a, 1981b; Reid et al., 1987; deCharms et al., 1998; Klein et al., 2000; Schnupp et al., 2001; Miller et al., 2002; Christianson et al., 2008) are very useful for characterizing auditory neurons, since they are easy to understand and give an intuitive idea of neurons' response properties. However, the auditory system is anatomically and physiologically very complex, as described in the previous section. Neurons in the auditory cortex are functionally much more complex than a linear filter, with various non-linearities (Linden et al., 2003; Atencio et al., 2008)

and context-dependent effects (Ahrens et al., 2008; Mesgarani et al., 2009; Williamson et al., 2016; David, 2018). Modelling approaches should therefore consider the inherent computational complexity of the auditory system.

One important aspect of any encoding model is that the model should be able to predict the response of a neuron to novel stimuli. When a linear STRF model is used to predict a neuron's response to novel stimuli, the prediction performance can be poor (Machens et al., 2004). An STRF is typically generated by regression or spike-triggered averaging for uncorrelated stimuli (Chichilnisky, 2001; Meyer et al., 2017). However, natural stimuli are correlated and so estimation of an STRF using natural stimuli requires removing this correlation from the STRF. Another way to estimate the STRF for natural stimuli is by using regularized linear regression. Regularization involves applying a penalty to STRFs which have implausible structure (Theunissen et al., 2000; Linden et al., 2003; Machens et al., 2004). For example, L1 regularization penalizes STRFs which have many non-zero coefficients and results in 'sparse' STRFs with relatively few non-zero coefficients. Regularization reduces overfitting and results in STRFs which tend to have better predictive performance when tested on novel data. More detail on the regularization techniques we used is in a following section - Training, testing, and analyzing encoding models. The prediction performance of STRFs can be increased by adding a static non-linearity to the linear model (Machens et al., 2004; Ahrens et al., 2008; Atencio et al., 2008; Calabrese et al., 2011; Rabinowitz et al., 2012; Williamson et al., 2016). Recently, more biologically inspired models, which have characteristics such as connectivity (Stevenson et al., 2012), network structure (Harper et al., 2016; Willmore et al., 2016), synaptic depression (David and Shamma, 2013) and adaptation (Espejo et al., 2019) have been shown to further improve prediction capacities somewhat. Besides aiming to improve the prediction accuracy of neural responses, modelling

approaches have been undertaken to improve our understanding of neurons' receptive fields (Thorson et al., 2015) or to identify the dimensionality of neural responses (Sharpee et al., 2004).

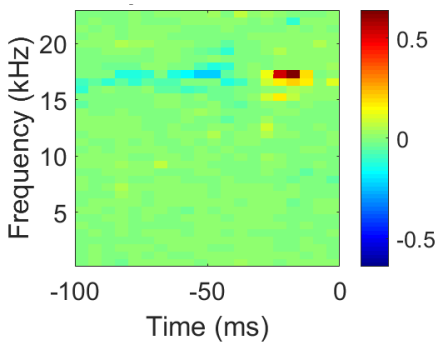

Figure 1.4: A spectrotemporal receptive field (STRF) of an auditory cortical neuron. *The x- and y-axes indicate the time relative to the present and sound frequency, respectively. In the colorbar, red indicates excitation and blue indicates inhibition by the stimulus. The neuron shown in the example is tuned to a high frequency and a relatively recent time in the past.*

1.5 Training, testing, and analyzing encoding models

A very general method to fitting encoding models is to use the maximum likelihood approach (Paninski, 2004; Meyer et al., 2017), in which an objective function is minimized between the prediction of a model and actual neural response. However, this approach is often prone to overfitting. Various methods exist however to prevent overfitting. Regularization (Friedman et al., 2010) and cross-validation are such methods. Regularization prevents overfitting by penalizing more complex models over simple ones (complex models are more prone to overfitting). By cross-validation, a model can first be trained and tested on a fraction of the dataset, and as a result the 'hyperparameter' governing the strength of the regularization can be chosen. After choosing the

regularization parameter, the model can then be tested on a held-out dataset. This ensures that the model is not biased to the training data and can generalize outside of the training examples. One benefit of using this method is that this can be used for a wide range of models, from linear models to linear non-linear models, complex deep and recurrent neural networks.

Because performance of a model can be assessed through its ability to predict, this approach also allows one to directly choose between different types of models. The premise is that the model with best prediction performance would be the model that represents the neural system from which the data are obtained. But prediction performance is not the only thing that matters. Understanding a system often means working out how an operation is performed by the system. In the case of cortical neurons, understanding the response properties of individual neurons or a group of neurons in the light of the underlying network can produce valuable insight. For example, this includes elucidating the biological mechanisms underlying an STRF. However, prediction can offer a very strong tool to determine which model, and thus which set of computations, best approximates the actual computations performed (Yamins and DiCarlo, 2016; Cadena et al., 2019; Richards et al., 2019).

1.6 Decoding neural responses

While encoding models predict the activity of individual neurons or of a neural population in a brain area, decoding functions do the reverse. For a given stimulus parameter, s, if a neuron produces the response, r, then function, D, which relates these two is a decoding function, that is $s = D(r)$.

Decoding functions are used to reconstruct the stimulus or values of a stimulus parameter from neural responses (Gielen et al., 1988; Hesselmans and Johannesma, 1989; Bialek et al., 1991). Decoding can provide valuable insights about the nature of information the nervous system encodes without the need for explicit assumptions about the nature of transformations in the neural circuits. Furthermore, the brain itself might be solving a decoding problem by estimating underlying stimulus parameters from sensory signals. So, studying decoding can provide us with important clues as to how the brain decodes information from the environment. In the auditory system, decoding has been applied to reconstruct synthetic and natural stimuli from neural responses (Mesgarani et al., 2009; Rabinowitz et al., 2013; Santoro et al., 2014; Yildiz et al., 2016), to study music (Thompson, 2013), to decode auditory attention (Mesgarani and Chang, 2012; Lauteslager et al., 2014; Wong et al., 2018) and to study speech processing (Ramirez et al., 2011; Ding and Simon, 2012; Mesgarani and Chang, 2012; Pasley et al., 2012; Martin et al., 2014; Mesgarani et al., 2014; Akram et al., 2016; O'Sullivan et al., 2017; Puvvada and Simon, 2017; Akbari et al., 2019). Our use of decoding models will be for mostly descriptive and interpretive purposes.

1.7 Experimental setup

We use ferrets as the model organism for this study. Ferrets are well-suited for the study of the auditory system because they can be readily trained in a variety of auditory tasks, their auditory cortex is readily accessible, and their audible frequency range overlaps with that of humans (Nodal and King, 2014). We study both anesthetized and awake animals although most of the thesis chapters will involve analysis of experimental data from anesthetized animals. Anesthetized animals have the advantage of lacking any motor activity or behavior so that the stimuli driven activity can be independent of factors such as movement, arousal, and attention. Neural response data were collected by extracellular electrophysiological recording using multielectrode probes.

In addition, a large part of the data that we analyze was obtained from other studies (Harper et al., 2016; Willmore et al., 2016; Espejo et al., 2019).

1.8 Summary of thesis

In the current thesis, we will investigate the encoding, and to a degree decoding, of natural sounds in the auditory system. This work is based on a collection of neural recordings from ferret auditory cortex in response to various natural and synthetic sounds. We aim to better understand the transformation of sound stimulus from the ear to the cortex, the mechanisms that underlie this transformation and the ability of a neural population to predict the future of ongoing sound stimuli. Central to this objective is the ability to build a model that will predict neural responses to a diverse selection of stimuli.

While previous modelling studies have provided significant advances in modelling the responses of the auditory cortex to natural sounds, currently available models are still very far from achieving 100% accuracy in predicting the responses of neurons. Furthermore, neurons in the auditory system are part of an intricate network of neighboring neurons with individual neuron having complex biological properties. We set out to improve the prediction performance of the model of auditory cortex and also to simultaneously investigate the computational role of various biological features along the auditory pathway. We will do this by exploring a wide range of models that differ in their level of biological details. Furthermore, we will also reconstruct sound stimuli from the activity of neural populations recorded in the auditory cortex. We will investigate the ability of the auditory cortex to estimate past inputs and to predict future inputs. Taken together, we hope this research will improve our understanding of how auditory system processes sounds to represent them as the spiking activity of neurons.

Chapter 2 | Suitable input for models of auditory cortical neurons

This chapter has been previously published as and adapted from: Rahman M, Willmore BDB, King AJ, Harper NS (2020) Simple transformations capture auditory input to cortex. PNAS 117 (45): 28442 – 28451

2.1 Introduction

Neural encoding models are used to explain and predict how neural responses depend on sensory stimuli. For such models, the raw stimulus may not be the ideal input to the model. Instead, a pre-processed version of the stimulus may be used as input. Such pre-processing might reflect, for example, the stimulus transformation occurring in the sensory periphery. In the auditory system, the sound pressure waveform is not a suitable input to an encoding model of neurons in the primary auditory cortex. A better choice of input is a frequency-decomposed version of the sound (Aertsen et al., 1980; Aertsen and Johannesma, 1981a; Eggermont et al., 1983; deCharms et al., 1998; Klein et al., 2000; Theunissen et al., 2000; Schnupp et al., 2001; Miller et al., 2002; Escabí and Schreiner, 2002; Linden et al., 2003; Fritz et al., 2003; Gill et al., 2006; Christianson et al., 2008; David et al., 2009; Gourévitch et al., 2009), resembling the peripheral pre-processing that takes place in the cochlea.

Sensory systems, from the sense organs up through the neural pathway, are typically very complex, comprising many different structures and cell types that often interact in a non-linear fashion. The complexity of these dynamic systems can make understanding their computations challenging. However, much of this physiological complexity may reflect biological constraints or come into play only under unusual conditions. Consequently, it could be that the signal

transformations that they commonly compute are substantially simpler than their physical implementations (Marr and Poggio, 1976). This can have implications for finding the appropriate model of the transformation of auditory signals through the ear to the cortex. To find out what peripheral model best captures such transformation, we appended various models of the auditory periphery to neural encoding models to predict auditory cortical responses to diverse sounds. We used both synthetic and natural sounds, as the latter are central to the normal function of the auditory pathway.

Various models of the auditory periphery have been developed and refined (Lyon, 1982, 2011; Wang and Shamma, 1994, 1995; Chi et al., 2005; Meddis, 2006; Meddis et al., 2013; Saremi and Stenfelt, 2013; Steadman and Sumner, 2018; Verhulst et al., 2018) to account for experimental observations of cochlear and auditory nerve properties in different species and for human psychophysical data. Some models are biologically detailed and accurately capture particular response properties of the auditory nerve (Meddis et al., 2001; Sumner et al., 2002, 2003; Meddis, 2006; Zilany et al., 2014; Bruce et al., 2018; Steadman and Sumner, 2018), while others are abstracted approximations of the signal transformation in the auditory periphery (Harper et al., 2016; Willmore et al., 2016; Rahman et al., 2019). Some have been used to provide inputs for models of auditory neurons (David et al., 2009; Carney et al., 2015; Harper et al., 2016; Willmore et al., 2016; Rahman et al., 2019) to generate perceptual models (McDermott and Simoncelli, 2011) and in machine processing of sounds (Lyon, 1982; Russo et al., 2019). However, few attempts (Thorson et al., 2015) have been made to determine which cochlear models best describe the computational impact of the auditory periphery on neural responses in mammalian auditory cortex, although more progress has been made in the avian auditory system (Gill et al., 2006). The models that best explain particular physiological characteristics of the auditory periphery may differ from the ones that best explain the impact of auditory nerve activity on cortical

responses to natural sounds. This is because neuronal responses are transformed through the central auditory pathway to the cortex, and the periphery may operate differently with natural sounds.

Here we considered a range of existing biologically detailed models of the auditory periphery, and adapted them to provide input for a number of encoding models of cortical responses. We also constructed a variety of simple spectrogram-based models, including a novel one accounting for the different types of auditory nerve fiber. Surprisingly, we found that the responses of neurons in the primary auditory cortex (A1) in ferrets can be explained equally well using the simple spectrogram-based cochlear models, as when more complex biologically-detailed cochlear models are used. Furthermore, the simple models explain the cortical responses more consistently well over different sound types and anesthetic states. Hence, much of the complexity present in auditory peripheral processing may not substantially impact cortical responses. This suggests that the intricate complexity of the cochlea and the central auditory pathway together results in a simpler than expected transformation of auditory inputs from ear to cortex.

2.2 Results

2.2.1 Generating cochleagrams using cochlear models

In this study, we consider two broad classes of cochlear models. The first class are based on cochlear filterbanks and have somewhat more detailed biological underpinnings than the second class. We consider several models of this first class, which we refer to here as the Wang Shamma Ru (WSR) model (Wang and Shamma, 1994, 1995; Chi et al., 2005), the Lyon model (Lyon, 1982, 2011), the Bruce Erfani Zilani (BEZ) (Zilany et al., 2009, 2014; Bruce et al., 2018) model and the Meddis Sumner Steadman (MSS) model (Meddis et al., 2001, 2013; Sumner et al., 2002; Meddis, 2006; Steadman and Sumner, 2018). These models vary substantially in their filterbanks and compression functions (see Methods for details). The WSR model has logarithmically-spaced filters, followed by non-linear compression, lateral inhibition and leaky integration (Ru, 2001). The Lyon model has a near log spacing of frequency channels that becomes more linear near the low frequencies. The frequency decomposition is accompanied by an adaptive gain control mechanism that acts as the compression function (Lyon, 1982, 2011). The BEZ model includes multiple detailed stages of signal transformation to mimic various stages of the processing by the ear and the auditory nerve of the cat (Zilany et al., 2009; Zilany and Carney, 2010; Bruce et al., 2018). The MSS model is similar to the BEZ model in that it also models the processing stages from the ear to the auditory nerve, but of a different species, the guinea pig (Meddis, 2006; Meddis et al., 2013) (Fig. 2.1). Recent work suggests the ferret peripheral auditory system is comparable to that of other mammalian species (Sumner et al., 2018), such as cats and particularly guinea pigs (Sumner and Palmer, 2012).

The second class of models are the STFT (short-time Fourier Transform) spectrogram-based models – these models are aimed at approximating the information processing in the auditory

periphery without modelling the detailed biological mechanisms (Harper et al., 2016; Schoppe et al., 2016; Willmore et al., 2016; Rahman et al., 2019). Implementation of these models consists of three key components: frequency decomposition, response integration and compression. We constructed the spectrogram-based cochlear models by performing frequency decomposition using a short-time Fourier transform of the sound waveform. The amplitude or power spectrogram was then put though a weighted summation using overlapping triangular filters spaced on a logarithmic scale to obtain specified numbers of frequency channels. Finally, non-linear compression was applied. For the amplitude spectrogram-based models, the compression functions used were a thresholded log function, and a log(1+(·)) function. We refer to these models as the spec-log and spec-log1plus, respectively. For the power spectrogram-based models, a thresholded log compression function was used either alone or together with a Hill function; we refer to these models as the spec-power and spec-Hill models (Fig. 2.1).

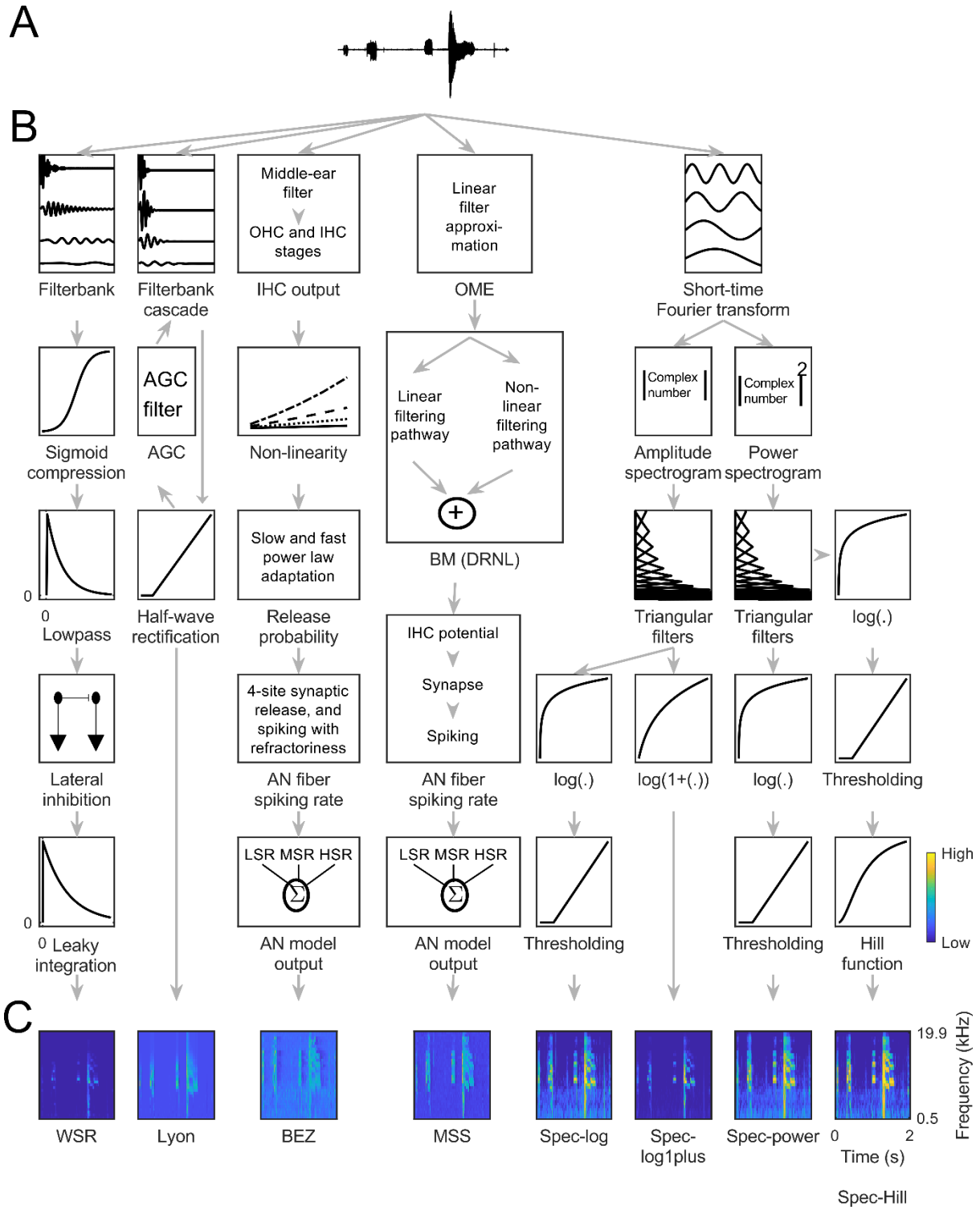

Figure 2.1: Schematics of the cochlear models. *A. A sound waveform, the input to a cochlear model. B. The stages of transformation of sound through each of the cochlear models (from left to right): Wang Shamma Ru (WSR) model (Wang and Shamma, 1994, 1995; Ru, 2001; Chi et al.,*

2005), Lyon model (Lyon, 1982, 2011), Bruce Erfani Zilani (BEZ) model (Zilany et al., 2009, 2014; Bruce et al., 2018), Meddis Sumner Steadman (MSS) model (Meddis et al., 2001, 2013; Sumner et al., 2002; Meddis, 2006; Steadman and Sumner, 2018), spec-log model, spec-log1plus model, spec-power model and spec-Hill model (see Methods). AGC, adaptive gain control; OME, outer and middle ear; OHC, outer hair cell; IHC, inner hair cell; BM, basilar membrane; DRNL, dual resonance non-linear filter; lin, linear; nonlin, nonlinear; AN, auditory nerve; LSR, low spontaneous rate; MSR, medium spontaneous rate; HSR, high spontaneous rate. C. The output of the cochlear models, the cochleagram. The example shown here is a 3 s excerpt of a sound of a wolf howling by a waterfall.

Each cochlear model produces a characteristic cochleagram for the same sound input. We illustrate this by presenting a range of synthetic and natural sound inputs (Fig. 2.2A) to each model. Figure 2.2B shows the cochlear models' responses to a click, pure tones of 1 kHz and 10 kHz, white noise and natural sounds. Here we depict cochleagrams with 32 frequency channels, although we studied the impact of varying the number of channels in each model, typically examining 2, 4, 8, 16, 32, 64 and 128 frequency channels. For a click input, cochleagrams produced by spectrogram-based models have sound activity tightly localized in time, but cochleagrams produced by filterbank-based models are more temporally spread, with the response persisting after the impulse occurred (Fig. 2.2B). For pure-tone input, cochleagrams produced by all models look similar except for the Lyon, BEZ and MSS models, where the cochleagram is broader in frequency content than in the other models (Fig. 2.2B). For white noise, most models have responses smoothly distributed across frequency, except for the WSR model.

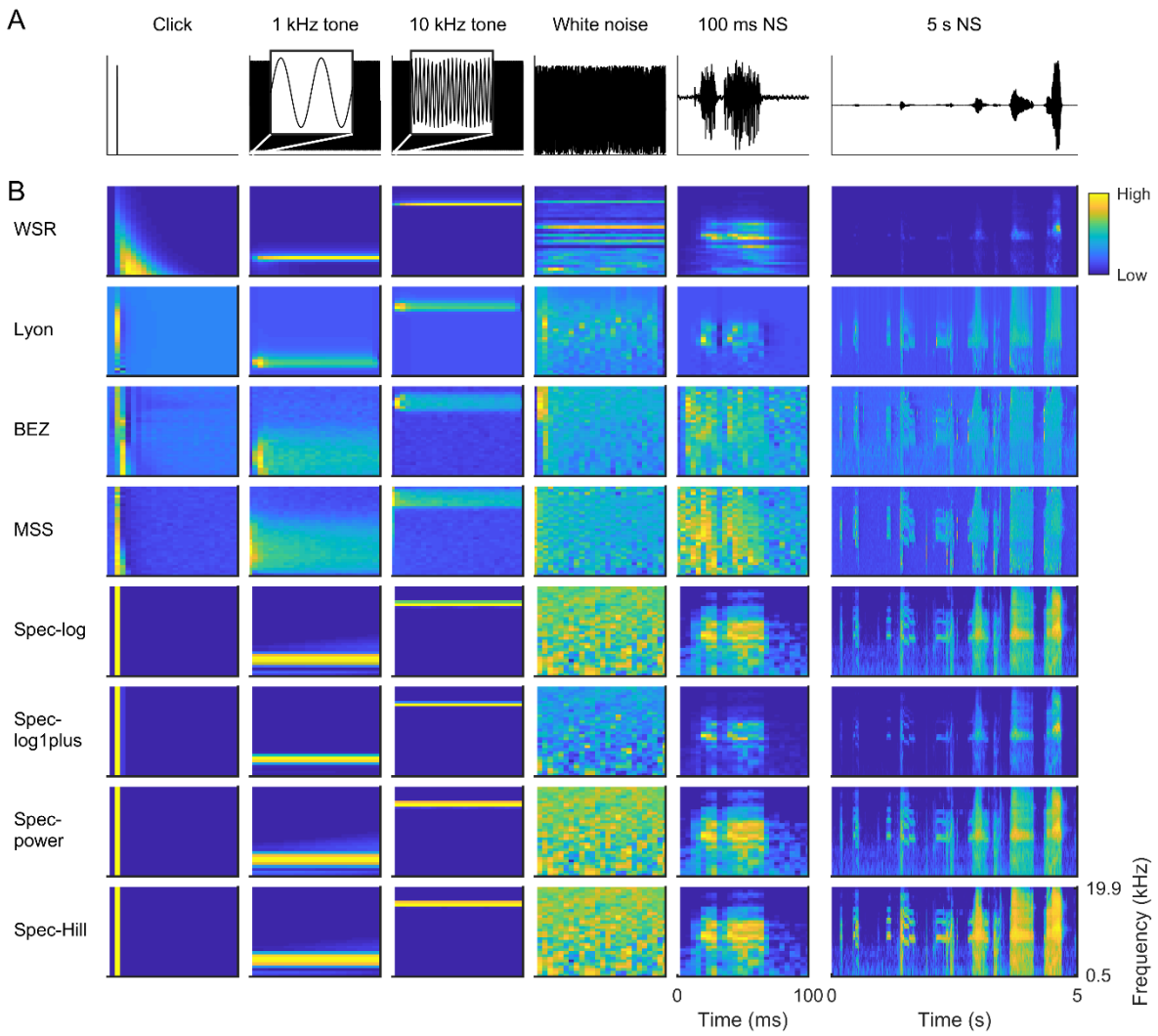

Figure 2.2: Cochleagram produced by each cochlear model for identical inputs. *A. Each column is a different stimulus: a click, 1 kHz pure tone, 10 kHz pure tone, white noise, a natural sound – a 100 ms clip of human speech and a 5 s clip of the same natural sound (from left to right). B. Each row is a different cochlear model.*

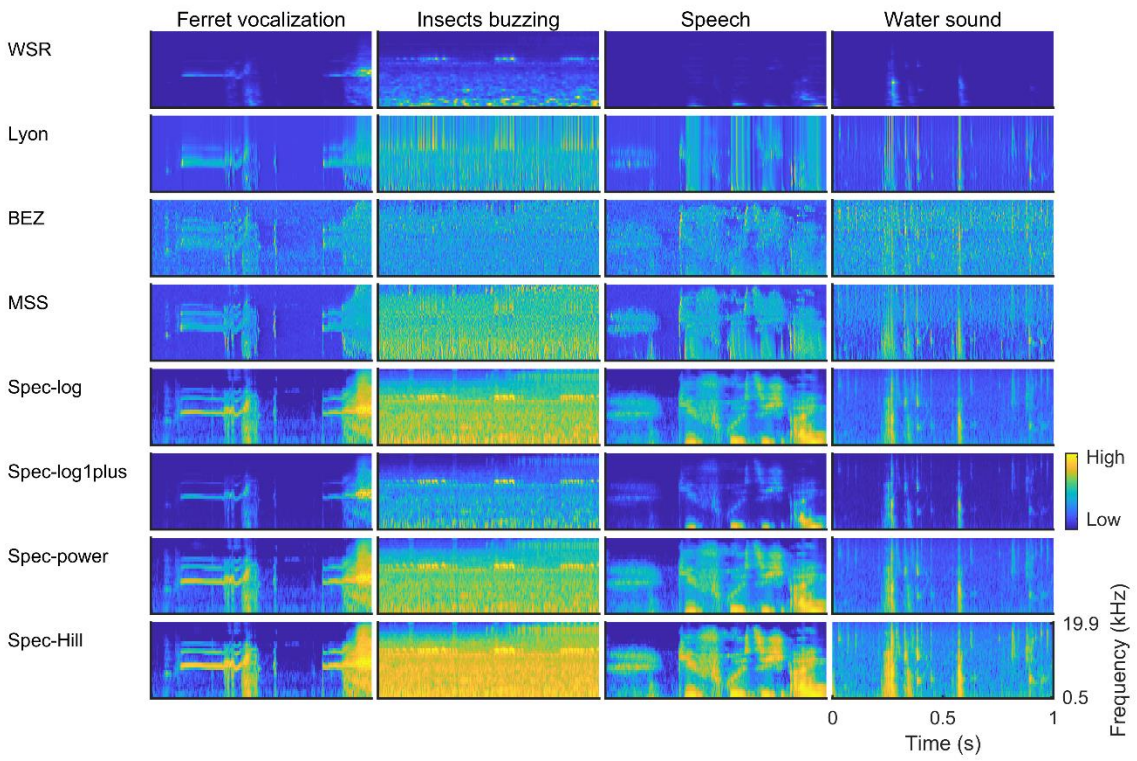

Figure 2.3: Examples of cochleagrams of natural sound stimuli. *Cochleagrams of 1 s excerpts of various natural sound clips are shown here. All cochleagrams contain 32 frequency channels. Only very high energy regions are visible in WSR cochleagrams since its compressive function does not convert sound energy to a range that is distributed evenly over the colormap.*

Cochleagrams of natural sounds also differ between cochlear models (Fig. 2.2, 2.3). However, two filterbank-based models (the BEZ and MSS model) produce similar looking cochleagrams, as do three spectrogram-based models (spec-log, spec-power and spec-Hill). Overall, the WSR model produced very different cochleagrams across a range of stimuli (click, white noise and natural sound). Furthermore, the maximum energy in the cochleagrams of the spec-log1plus model is lower than other spectrogram-based models. Compared to other models, the output of the BEZ and MSS models looks noisier due to the stochasticity in their models of inner hair cells, ribbon synapse vesicle release or auditory nerve firing. We therefore averaged across multiple repeated runs of these models to provide cochleagrams that reduced this variability (see Methods for more details). We quantified the similarity between the cochleagrams produced by each cochlear model for natural sound inputs by calculating the correlation coefficients between the cochleagrams produced by each possible pair of cochlear models. This quantitative analysis supports our qualitative observations of the similarities and differences between the cochleagrams of the different models (Fig. 2.4).

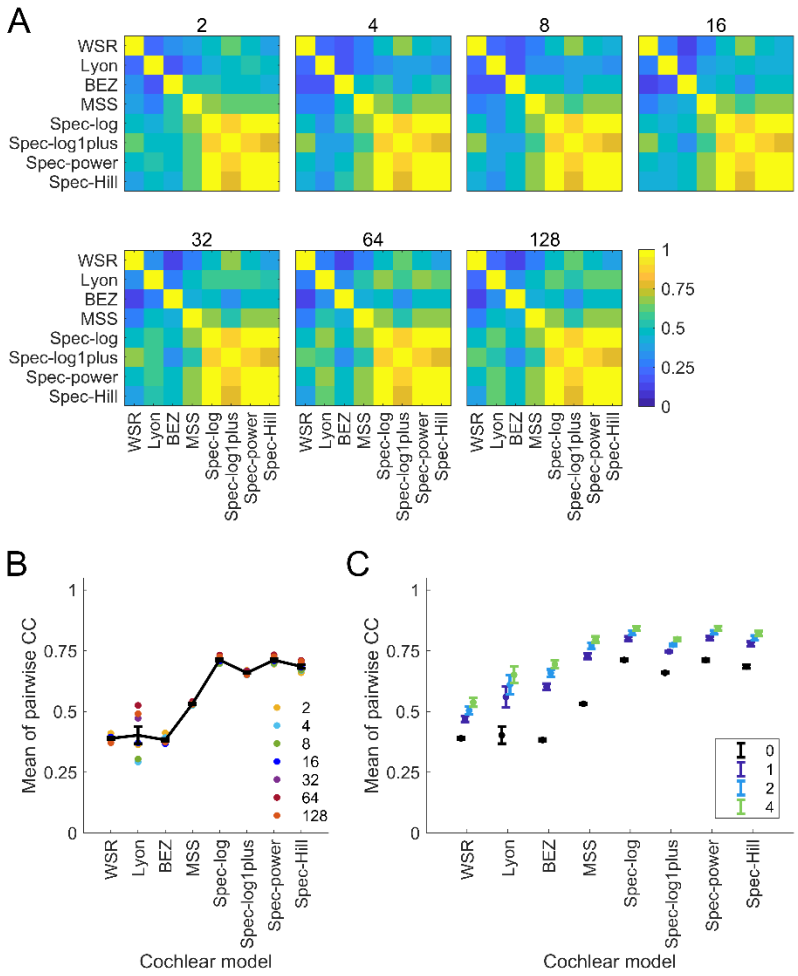

Figure 2.4: Comparison of cochleagrams produced by different models. *A. Correlation coefficient (CC) between cochleagrams produced by all possible pairs of models. The number on top of each plot shows the number of frequency channels in the cochleagram. B. Mean of the pair-wise CC of a model with the rest of the models. Colored dots show the mean for a model with a specified number of frequency channels and the black line with error bars show the mean over the means for different frequency channel numbers. C. Mean of the means of the pair-wise CCs after introducing Gaussian blurring to each cochleagram to account for spectral and temporal shift.*

2.2.2 Predicting responses of auditory cortical neurons using different cochleagram inputs

The datasets used in this study were from extracellular recordings of the responses to sounds of neurons in ferret A1. We used three datasets: responses to natural sounds in anesthetized ferrets (natural sound dataset 1 – NS1), responses to dynamic random chords in the same anesthetized ferrets (DRC dataset), and responses to natural sounds in awake ferrets (natural sound dataset 2 – NS2). We will focus first on the results with natural sound dataset 1, which consisted of neural responses to a diverse selection of natural sounds (20 sound snippets, each 5s in duration), including human speech, animal vocalizations and environmental sounds (Harper et al., 2016; Willmore et al., 2016; Rahman et al., 2019). This dataset constitutes a total of 73 single units, which were those units with a noise ratio (Linden et al., 2003; Rabinowitz et al., 2011) of < 40 so as to exclude neurons whose response showed little dependence on the stimulus (see Methods and (Rahman et al., 2019) for details).

Cochlear models are often used as input to models of responses of auditory cortical neurons (Gill et al., 2006; Calabrese et al., 2011; Rabinowitz et al., 2011; Harper et al., 2016; Willmore et al., 2016; Rahman et al., 2019), such as the commonly used linear-nonlinear (LN) model of neural responses (Atencio et al., 2008; Rabinowitz et al., 2011). Hence, we use a two-stage encoding framework to estimate firing-rate time-series in response to natural sounds of neurons in ferret A1. The first stage of the encoding framework processes the sound stimuli using a cochlear model to generate a cochleagram (Fig. 2.5). The second stage estimates the firing-rate time-series as a function of the preceding cochleagram using an LN model (see Methods). An LN model was fitted individually to each unit's responses to 16 of the 20 sound snippets using k-fold (k=8) cross-validation and L_1-regularization (Friedman et al., 2010) (distribution of the values of the

regularization parameter is shown in Fig. 2.6). Details of the cross-validation procedure and parameter estimation have been described previously (Rahman et al., 2019) (also see Methods). We fitted an LN model for each cochlear model with a specified number of frequency channels (2, 4, 8, 16, 32, 64 and 128).

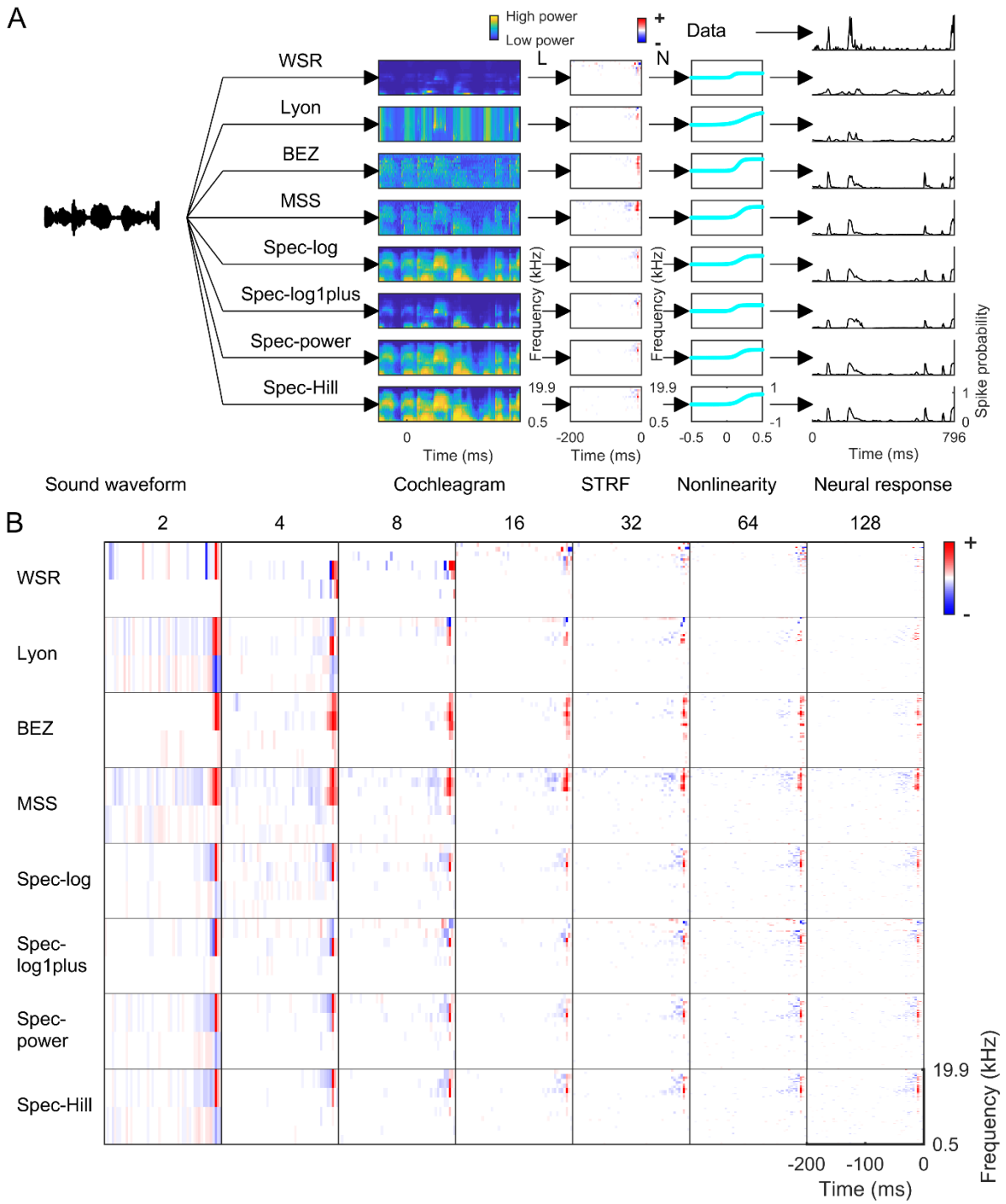

Figure 2.5: Estimating spectro-temporal receptive fields. *A. The encoding scheme: pre-processing by cochlear models to produce a cochleagram (in this case, with 16 frequency channels) followed by the linear (L)-nonlinear (N) encoding model. The parameters of the linear stage (the weight matrix) are commonly referred to as the spectro-temporal receptive field (STRF)*

of the neuron. Note how the choice of cochlear model influences estimation of the parameters of both the L and N stages of the encoding scheme and, in turn, prediction of neural responses by the model. B. The STRF of an example neuron from natural sound dataset 1, estimated by using different cochlear models. Each row is for a cochlear model and each column is the number of frequency channels.

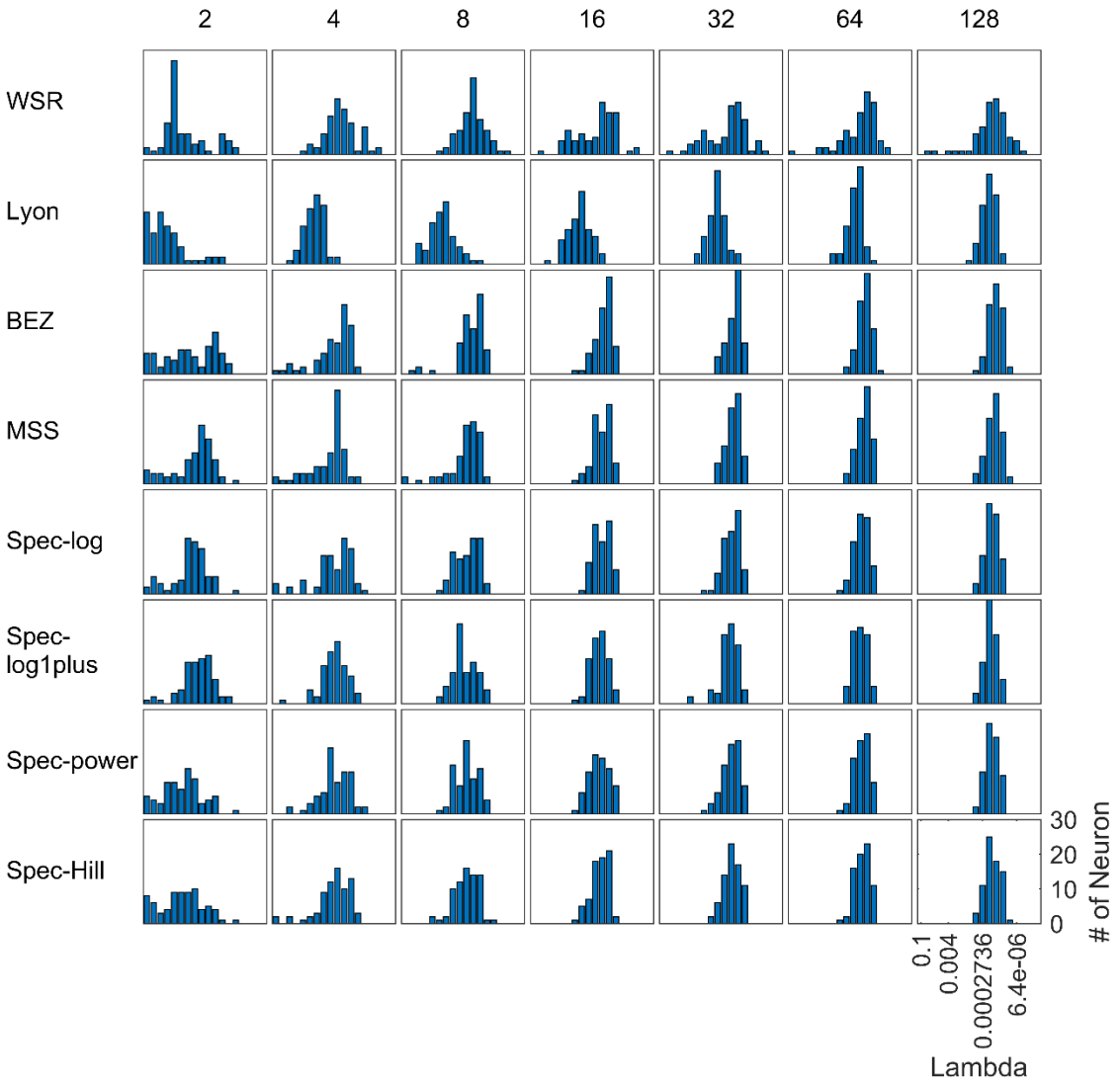

Figure 2.6: Distribution of the values of hyperparameter lambda in the LN model across the cortical neurons. *Each row shows the cochlear model that was used to generate the input for the LN model and each column shows the number of frequency channels in the cochleagram input. We explored a wide range of values for λ and the optimal λ values that were chosen to test the models are within the explored range, and do not typically lie at the edge of the range. This ensures that the variation in model performance is not due to poor choice of hyperparameter range.*

The linear part of the LN model captures the linear dependence of a neuron's firing rate on the frequency content of the cochleagram at different time delays, i.e. the spectro-temporal receptive field (STRF) (Aertsen et al., 1980; Aertsen and Johannesma, 1981a; deCharms et al., 1998; Theunissen et al., 2000; Klein et al., 2000; Schnupp et al., 2001; Miller et al., 2002; Escabí and Schreiner, 2002; Linden et al., 2003; Fritz et al., 2003; Gill et al., 2006; Christianson et al., 2008; David et al., 2009; Gourévitch et al., 2009). STRFs are widely used to describe the stimulus feature selectivity of auditory cortical neurons. The general properties of STRFs estimated for the same neuron using different cochlear models were similar (Fig. 2.5B). All cochlear models produced STRFs that contained excitatory and lagging inhibitory fields. The shape of the STRFs produced by different models also resembled each other. The largest weight in the STRF occurred at a comparable frequency (best frequency) and time (latency) for all models and regardless of the number of frequency channels. The only exception to this were the 2 and 4 frequency channel models, which sometimes showed very different frequency selectivity, presumably because of the very limited choice of frequency channels (Fig. 2.7). The ratio of inhibitory vs excitatory field strength (IE score) (Harper et al., 2016; Rahman et al., 2019) was also very similar for STRFs produced by different cochlear models, with the exceptions of the WSR and BEZ models (Fig. 2.7).

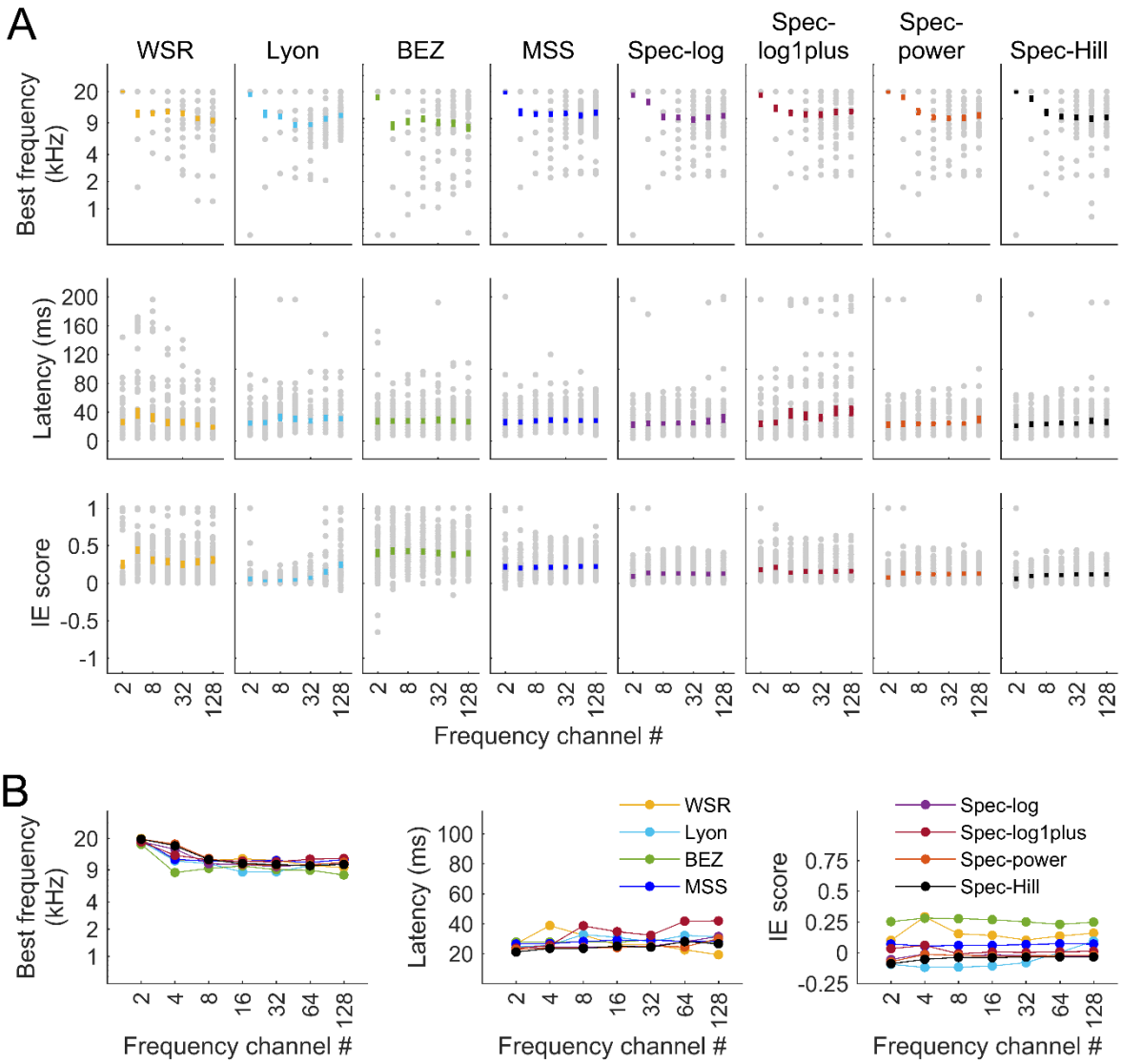

Figure 2.7: Cortical response properties estimated by using different cochleagram inputs to the LN model. *A. Best frequency, latency and IE score are shown in the top, middle and bottom rows, respectively, with each plot showing the estimated values using the model specified at the top and the number of frequency channels specified on the x-axis. Each grey dot is one neuron and the color bar indicates the mean and the squared error of the mean. B. Mean best frequency, latency and IE score over all neurons for each model and number of frequency channels.*

Although the general properties of the STRFs obtained using different cochlear models were similar, a more detailed analysis revealed some variability between pairs of STRFs estimated for the same neuron using two different cochleagram models (Fig. 2.8A-B). Higher correlations were observed between STRFs estimated from same class of cochlear models. In particular, spec-log, spec-power, and spec-Hill models produced very similar STRFs, whereas this was less true of the spec-log1plus model. STRFs obtained with the MSS and BEZ models were similar to each other and, to a lesser extent, to the Spec-Hill model. Applying Gaussian blurring to account for frequency or temporal shifts in the STRFs improves the correlations, but did not change the overall trends in these results (Fig. 2.8C).

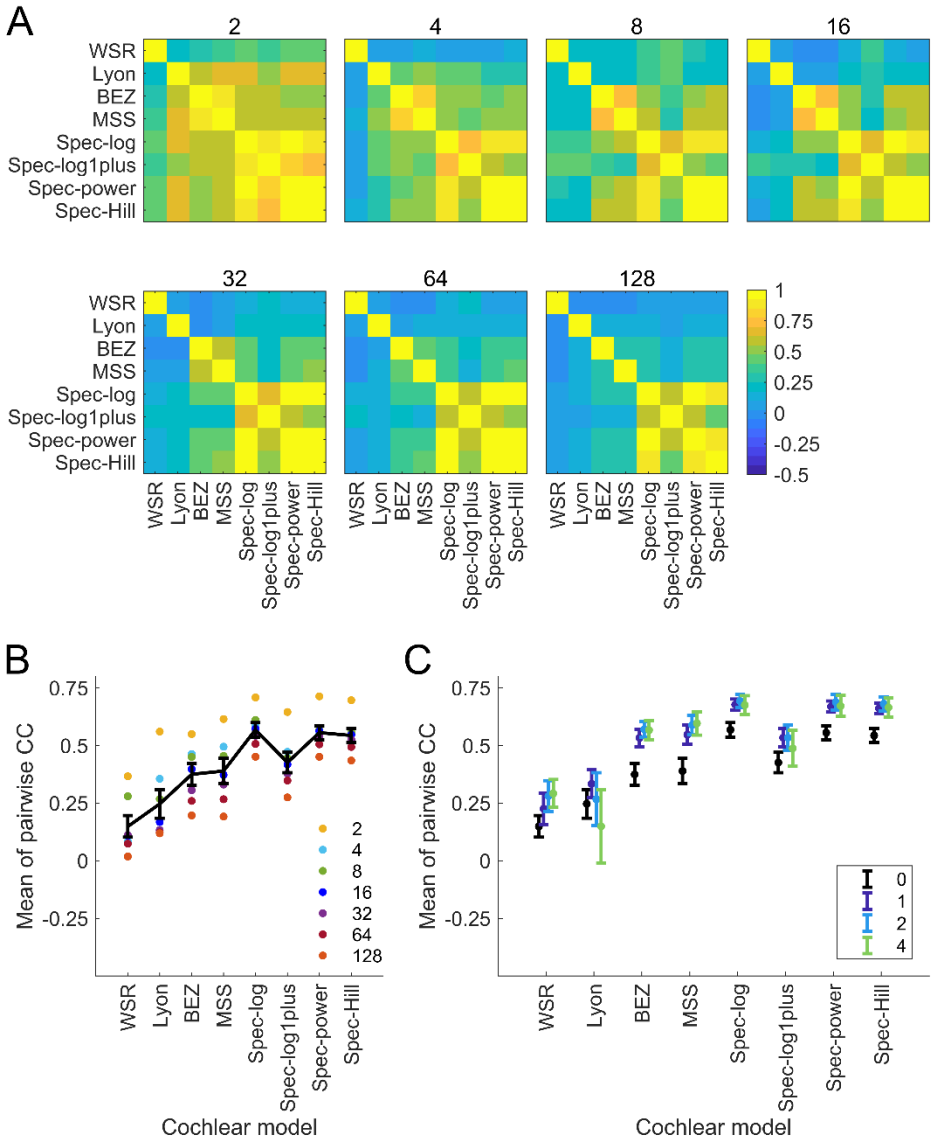

Figure 2.8: Comparison of STRFs estimated using with different cochleagram inputs. *A. Correlation coefficient (CC) between STRFs produced by all possible pairs of cochlear models. The number on top of each plot shows the number of frequency channels in the cochleagram. B. For each model, the mean CC of its STRFs with the STRFs of the other models with the same number of frequency channels. Colored dots show the mean for a model with a specified number of frequency channels and the black line with error bars shows the mean over the means for different numbers of frequency channels. C. Mean of the means of the pair-wise CCs after introducing Gaussian blurring to each STRF to account for spectral and temporal shift.*

The models vary in how well they match the peaks in the responses of A1 neurons. Likewise, the measured overall prediction performance of the LN model on a held-out dataset differed between different cochlear models. As a measure of the prediction performance we used the normalized correlation coefficient (CC_{norm}) (Schoppe et al., 2016) over all neurons in the dataset, where a CC_{norm} of 0 indicates no correlation between the neural response and the model's estimate and a CC_{norm} of 1 indicates that the model can predict all variance in the firing rate (averaged over repeats) that depends on the stimulus. We found that the mean CC_{norm} over all neurons varied depending on the choice of cochlear model and the number of frequency channels in the cochleagram (Fig. 2.9 and Supplementary Table 2.1). We define the peak CC_{norm} of a model as the highest mean CC_{norm} (averaged across all neurons) across the number of frequency channels. The peak CC_{norm} was 0.462 for the WSR model (at 8 channels), 0.662 for the Lyon model (at 64 channels), 0.644 for the BEZ model (at 128 channels), 0.725 for the MSS model (at 128 channels), 0.721 for the spec-log model (at 64 channels), 0.630 for the spec-log1plus model (at 64 channels), 0.722 for the spec-power model (at 64 channels) and 0.726 for the spec-Hill model (at 64 channels) (Fig. 2.9I and Supplementary Table 2.1). A log-spaced power spectrogram with successive log and Hill compression functions (spec-Hill) provided the best prediction performance, with a mean CC_{norm} of 0.726 (Supplementary Table 2.1). However, three of the spectrogram models, the spec-log, spec-Hill, spec-power, and one of the biological models, MSS, all predicted similarly well at about 0.72-0.73 peak CC_{norm}. In contrast, one of the spectrogram models, spec-log1plus, and three of the biological models, WSR, Lyon, and BEZ predicted substantially less well, with peak CC_{norm} in the range 0.45-0.66.

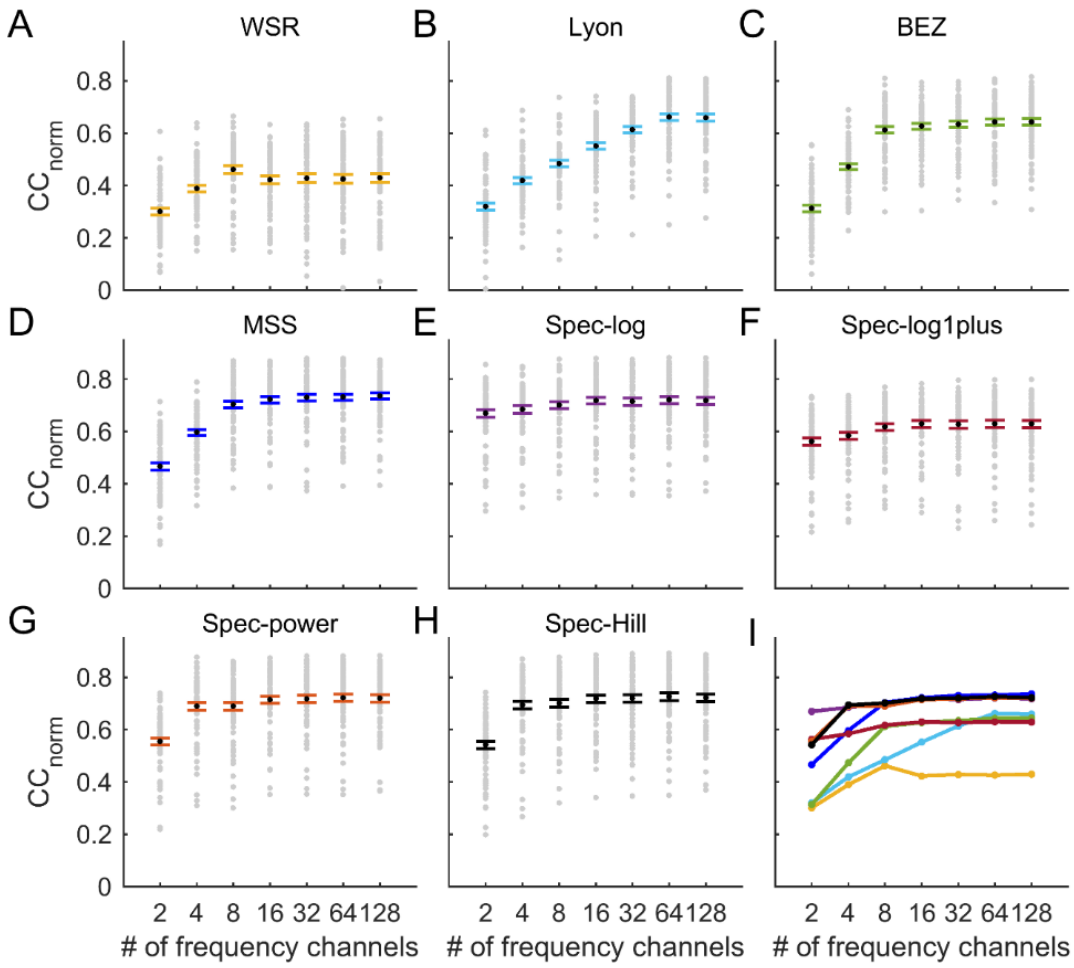

Figure 2.9: Performance of different cochlear models in predicting neural responses of natural sound dataset 1. *A. WSR model. B. Lyon model. C. BEZ model. D. MSS model. E. Spec-log model. F. Spec-log1plus model. G. Spec-power model. H. Spec-Hill model. Each grey dot represents the* CC_{norm} *between a neuron's recorded response and the prediction by the model; the larger black dot represents the mean value across neurons and the error bars are standard error of the mean. I. Comparison of all models. Color coding of the lines matches the other panels.*

Selecting the best model for individual neurons supports the findings based on the average performance of each model for all neurons. The spec-Hill and MSS models with either 64 or 128 frequency channels provided the best prediction performance for most neurons (Fig. 2.10). We also compared the predicted response obtained with the MSS model to the predicted response of the other models and found that the similarity in peak CC_{norm} performance generally co-varied with the similarity in predicted response (Fig. 2.11).

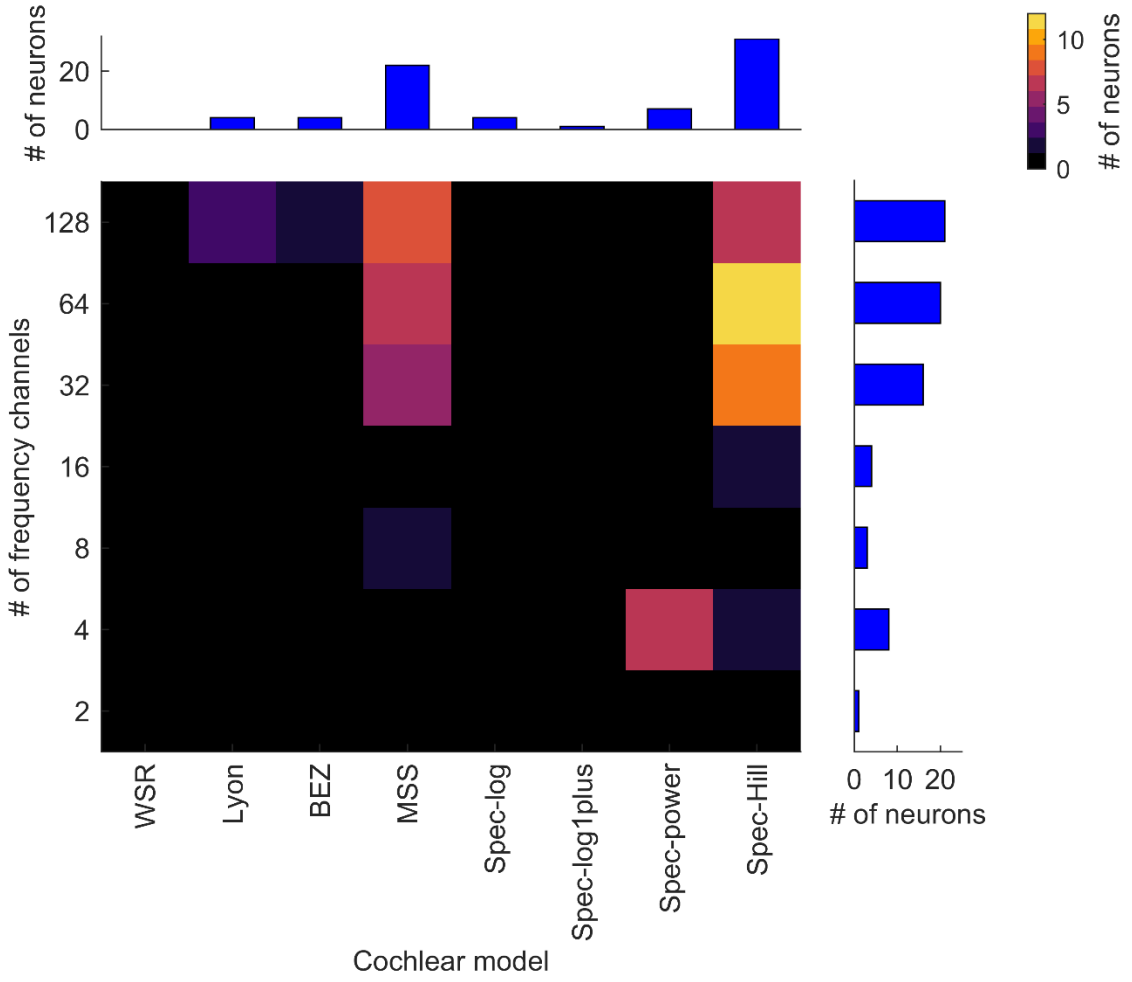

Figure 2.10: Best performing cochlear models for 73 primary auditory cortical neurons. *The heatmap shows the number of neurons for which each model with a specific number of frequency channels is the best performing model and the bars show the data collapsed along each axis.*

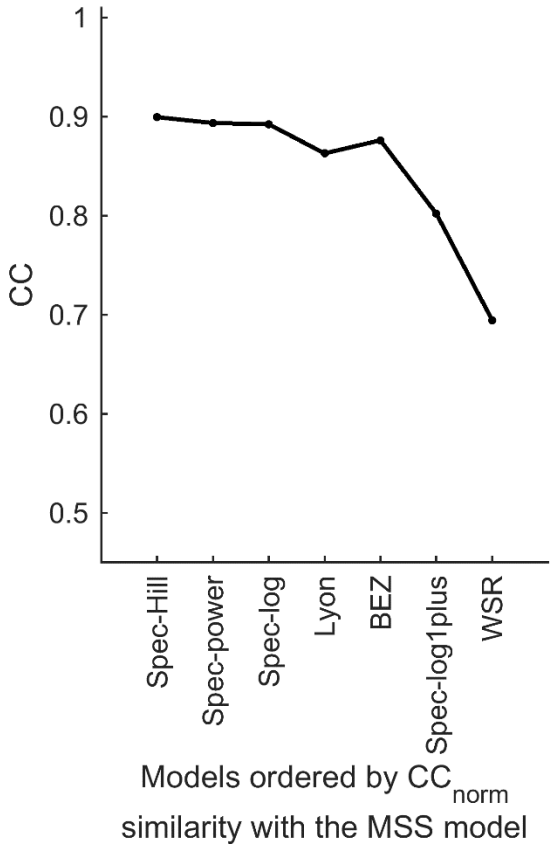

Figure 2.11: Similarity of the predicted response of the MSS model to the predicted response of other models. *Each dot represents the correlation co-efficient between the response of the MSS model and the specified model in the x-axis. Models on the x-axis are ordered by similarity of their* CC_{norm} *performance to show the relationship between prediction performance and response similarity.*

2.2.3 Multi-fiber cochleagrams

We have used the word cochleagram so far to refer to a time-frequency representation of the sound stimulus, with a single output for each time and frequency. In the auditory system, however, afferent nerve fibers tuned to the same frequency can have different sound intensity thresholds and dynamic ranges, and three different auditory nerve fiber types have been physiologically characterized (Sachs and Abbas, 1978; Colburn et al., 2003; Sumner and Palmer, 2012). The three types of fibers are low spontaneous rate (LSR) with higher threshold and larger dynamic range, medium spontaneous rate (MSR) with intermediate threshold and dynamic range, and high spontaneous rate (HSR) with lower threshold and narrower dynamic range. To study the impact of this representation on the prediction performance of modeled cortical responses, we used an MSS model with the three different fiber types (multi-fiber MSS model) as input to the LN model. We also constructed a multi-threshold spec-Hill model, where each frequency channel went through three different Hill functions with different thresholds and dynamic ranges (see Methods). This produced a cochleagram representation that assigns the changing sound level in a single frequency channel to three separate channels, analogous to three fiber types in the multi-fiber MSS model (Fig. 2.12A,B).

When we use these models as input to the LN model of cortical neurons, we were able to predict cortical responses to natural sounds slightly better than the single-fiber or single-threshold versions of the models (Fig 2.12). For the NS1 dataset, the multi-threshold spec-Hill model performed better than the multi-fiber MSS model for cochleagram inputs with fewer than 32 center frequencies, but performs slightly worse for cochleagram inputs with 32 or more center frequencies (Fig. 2.12). Detailed values of mean CC_{norm} are given in Supplementary Table 2.1.

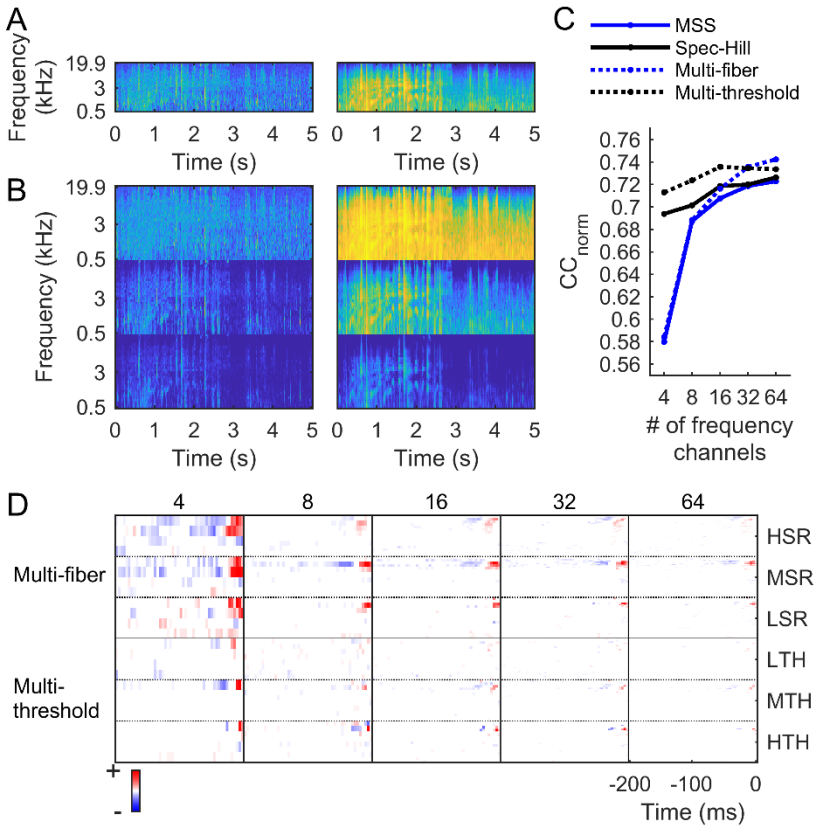

Figure 2.12: Multi-fiber and multi-threshold cochlear models. *A. Cochleagram of a natural sound clip produced by the MSS model (left) and the Spec-Hill model (right). B. Cochleagram of the same natural sound clip produced by the multi-fiber MSS model (left) and the multi-threshold Spec-Hill model (right). C. Mean* CC_{norm} *for predicting the responses of all 73 cortical neurons in natural sound dataset 1 for the multi-fiber/threshold models and their single-fiber/threshold equivalents. D. STRFs of an example neuron from natural sound dataset 1, when estimated using the multi-fiber and multi-threshold models.*

2.2.4 Generality of the model performance and further explorations

While we aimed with natural sound dataset 1 to have a diverse and representative stimulus, this does not, of course, represent the full space of natural sounds. Moreover, our electrophysiology data came from anesthetized animals, raising a question over whether brain state might affect model performance. To examine how the choice of stimulus and brain state influence the results, we tested the performance of the models on two other datasets. One of these (natural sound dataset 2 – NS2) consisted of extracellular recordings from A1 of awake ferrets in response to a different set of natural sounds (18 sound snippets, each 4 s in duration), including human speech, animal vocalizations, music and environmental sounds (Espejo et al., 2019). In total, 235 single units were included, which had a noise ratio <40. Using the same methods as for natural sound dataset 1, we used this new dataset to train and test the models' performance (see Methods). For this dataset, CC_{norm} values were lower for all models (Fig. 2.13A-B; Supplementary Table 2.1). Considering the peak CC_{norm} values, we found that the same simple spectrogram-based models (spec-log/power/Hill) remained among the top performing models, performing similarly to the best biological models (Lyon/BEZ) (Fig. 2.13A), at 0.33-0.34 peak CC_{norm}. However, the best-performing biological models were not the same as for natural sound dataset 1, with the MSS model now performing poorly compared to the Lyon and BEZ models (Supplementary Table 2.1). These worse performing models (spec-log1plus, MSS, and WSR) had peak CC_{norm} values within the range 0.22-0.32 (Supplementary Table 2.1).

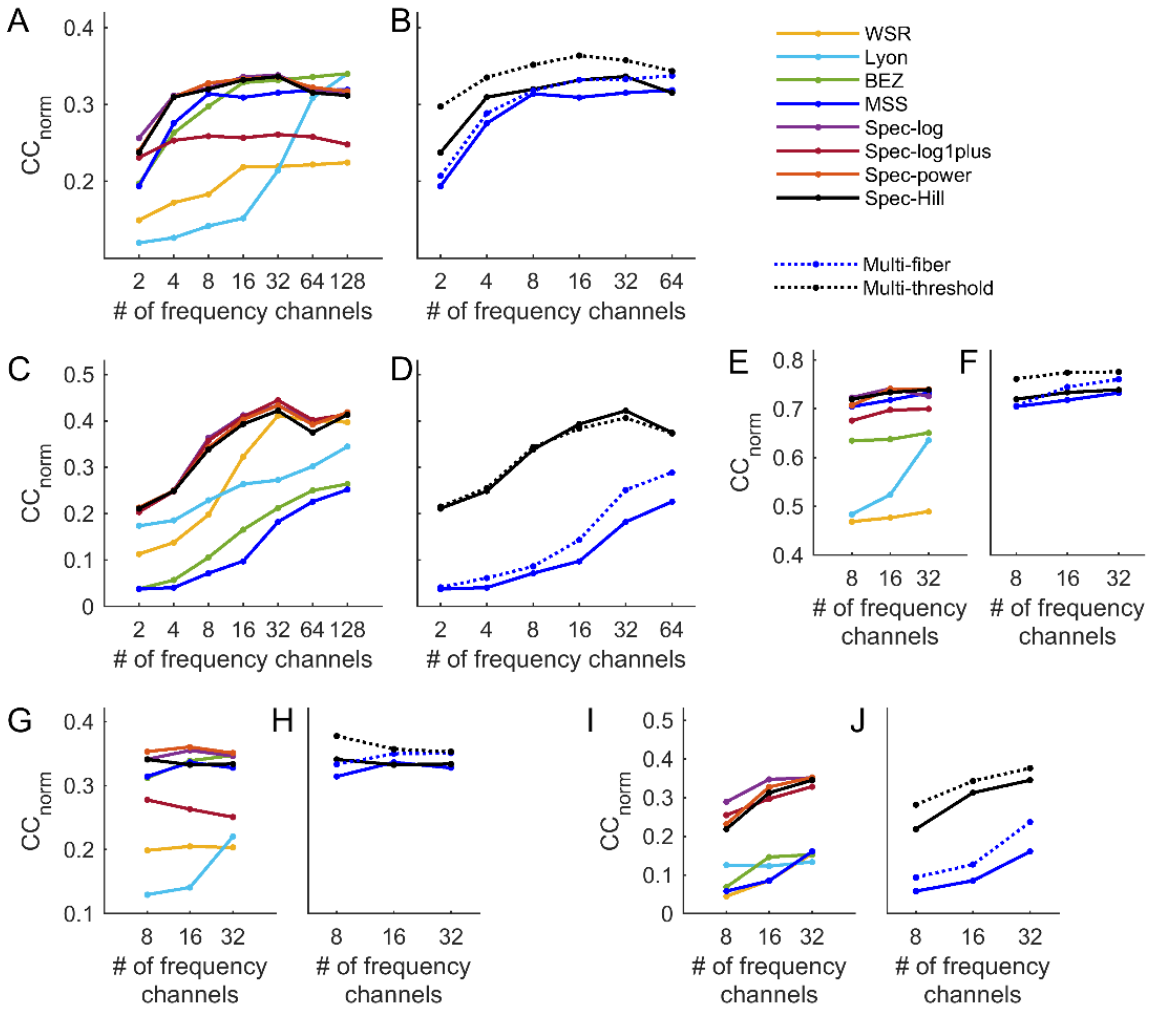

Figure 2.13: Performance of different cochlear models across datasets and encoding models. *A,B. Mean CC_{norm} between the LN encoding model prediction and actual data for all neurons in natural sound dataset 2 (awake ferrets) for single fiber models (A) and for multi-fiber models (B). C,D. Mean CC_{norm} between the LN encoding model prediction and actual data for all neurons in the DRC dataset (anesthetized ferrets) for single fiber models (C) and for multi-fiber models (D). E,F. Mean CC_{norm} between the prediction of the NRF model and actual data for all neurons in natural sound dataset 1 (anesthetized ferrets) for single fiber models (E) and for multi-fiber models (F). G,H. Mean CC_{norm} between the prediction of the NRF model and actual data for all neurons in natural sound dataset 2 for single fiber models (G) and for multi-fiber models (H).*

I,J. Mean CC_{norm} between the prediction of the NRF model and actual data for all neurons in DRC dataset for single fiber models (I) and for multi-fiber models (J).

We also tested the performance of the models on a different type of stimulus. The 73 neurons in natural sound dataset 1 were also played 12 dynamic random chord (DRC) stimuli, which consisted of randomly constructed chords changing every 25 ms. Using the same methods as for natural sound dataset 1, we used this DRC dataset to train and test the models' performances (see Methods). We found that the same simple spectrogram-based models (spec-log/power/Hill) remained among the top performing models, now joined by the spec-log1plus model (Fig. 2.13C-D; Supplementary Table 2.1). They performed slightly better than the best biological model (WSR) (Fig. 2.13C; Supplementary Table 2.1), with peak CC_{norm} in the range 0.42-0.45 compared to the 0.41 peak CC_{norm} of the WSR model. However, the biological models changed in which ones performed best, now with the MSS model (best for natural sound dataset 1) and Lyon/BEZ models (best for natural sound dataset 2) no longer being comparable to the spectrogram models, and instead the WSR model resembling the performance of the simpler spectrogram models. These worse performing models (MSS, BEZ and Lyon) had peak CC_{norm} values in the range 0.25-0.34 (Fig 2.13C; Supplementary Table 2.1). Thus, while the spec-log/Hill/power models show consistently high performances for all three datasets, the performance of the other more biological models varies substantially from dataset to dataset.

We found that for NS2 and DRC datasets the multi-threshold model outperformed the multi-fiber model. For NS2 dataset, we also found that both the multi-fiber model and multi-threshold model performed better than their single fiber/threshold equivalent models. However, for the DRC

dataset, while the multi-fiber model performs better than its single fiber equivalent (MSS) (Fig 2.13B), the multi-threshold model does not perform better than its single threshold equivalent (spec-Hill) (Fig. 2.13D). Detailed values of the prediction performance metric for all models and these two additional datasets are given in Supplementary Table 2.1.

So far, we have reported the prediction performance of the cochlear models when combined with an LN encoding model. To what extent does the choice of encoding model influence the results? We tested the prediction performance of just the linear stage (the STRF) of the LN model and found that CC_{norm} values are lower than those for the LN model. However, the performance of the models remain largely unchanged in comparison to one another (Fig. 2.14). Furthermore, we tested the performance of each cochlear model when combined with a network receptive field (NRF) encoding model (see Methods for more details) (17, 19). This is a single hidden layer neural network with units with sigmoid nonlinear activation functions. The NRF model has a higher number of parameters and is hence likely more sensitive to the amount of training data than the LN model. To keep the parameter number low and to save the running time, we ran these models with a limited set of frequency channel numbers. The CC_{norm} values for the NRF model for both natural sound datasets are typically higher than for the LN model. The NRF model values for natural sounds are particularly high when the NRF model is combined with the multi-threshold model (Figure 2.13F, H, I; Supplementary Table 2.2), reaching up to 0.78 for natural sound dataset 1. However, the relative performance of different cochlear models remains similar to the LN model (Fig. 2.13E-J; Supplementary Table 2.2).

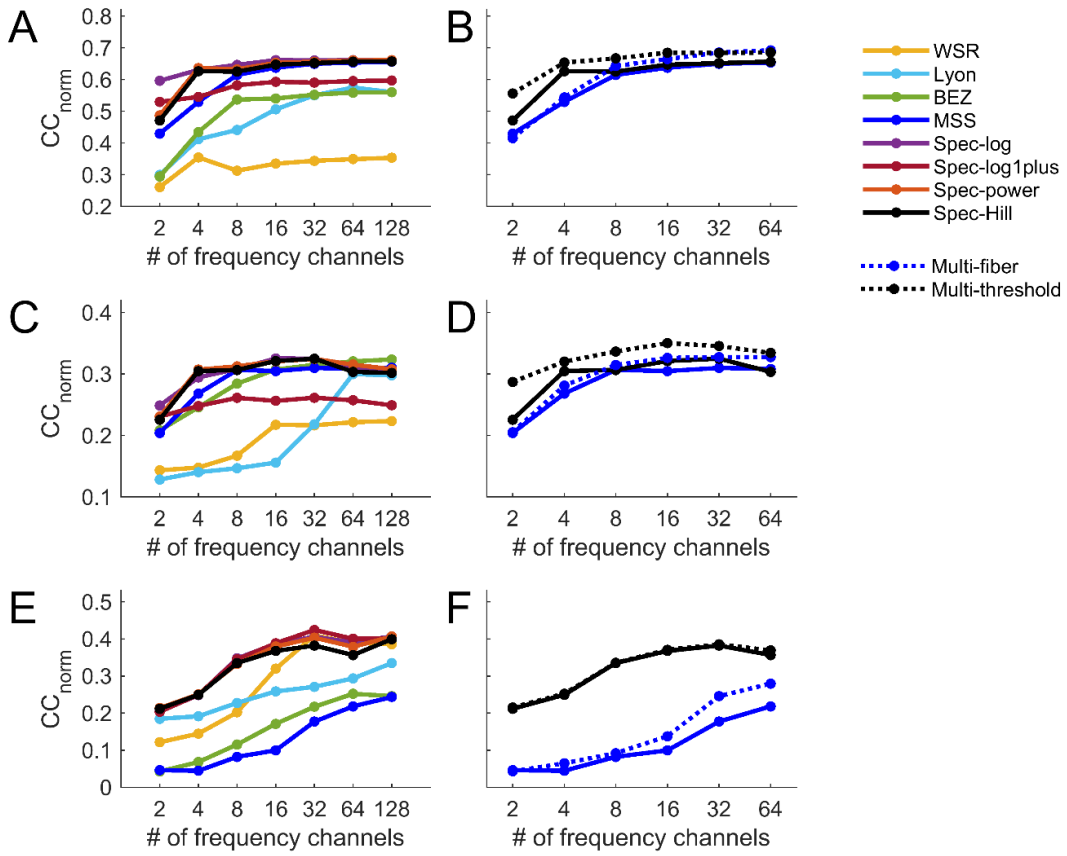

Figure 2.14: Average CCnorm for output of linear stage of the LN encoding model for all three datasets. *A and B. CCnorm on the natural sound dataset 1. C and D. CCnorm on the natural sound dataset 2. E and F. CCnorm on the DRC dataset.*

For all datasets, we examined the consequence of using CC_{norm}, by examining how the results looked for a different commonly used measure; the raw correlation coefficient (CC). The CC varies across the different models in a very similar way to the CC_{norm} (Fig. 2.15). We also explored in more detail the consequences of some of our other modelling choices, this time just for natural sound dataset 1. First, there is a stochastic element to the MSS and BEZ models. We investigated the effect of this noise in the MSS and BEZ model on the prediction performance. For predicting

cortical responses, we used the MSS and BEZ response averaged over 20 repeats to lessen the stochasticity. When we take the average of 100 or 200 repeats, the CC_{norm} of the MSS model is very similar to the MSS with 20 repeats, indicating that averaging over 20 repeats is sufficient (Fig. 2.16). The effect of repeats is also similar for the BEZ model (Fig. 2.17). Second, for model training and testing, we initially excluded onset responses from the neural data (the first 800 ms), as is common practice in STRF estimation (Depireux et al., 2001; Harper et al., 2016; Willmore et al., 2016). However, we have examined the consequence of including the onset responses, and we found that including them has very little effect on the performance of the LN models, regardless of the cochlear model used for preprocessing (Fig. 2.18).

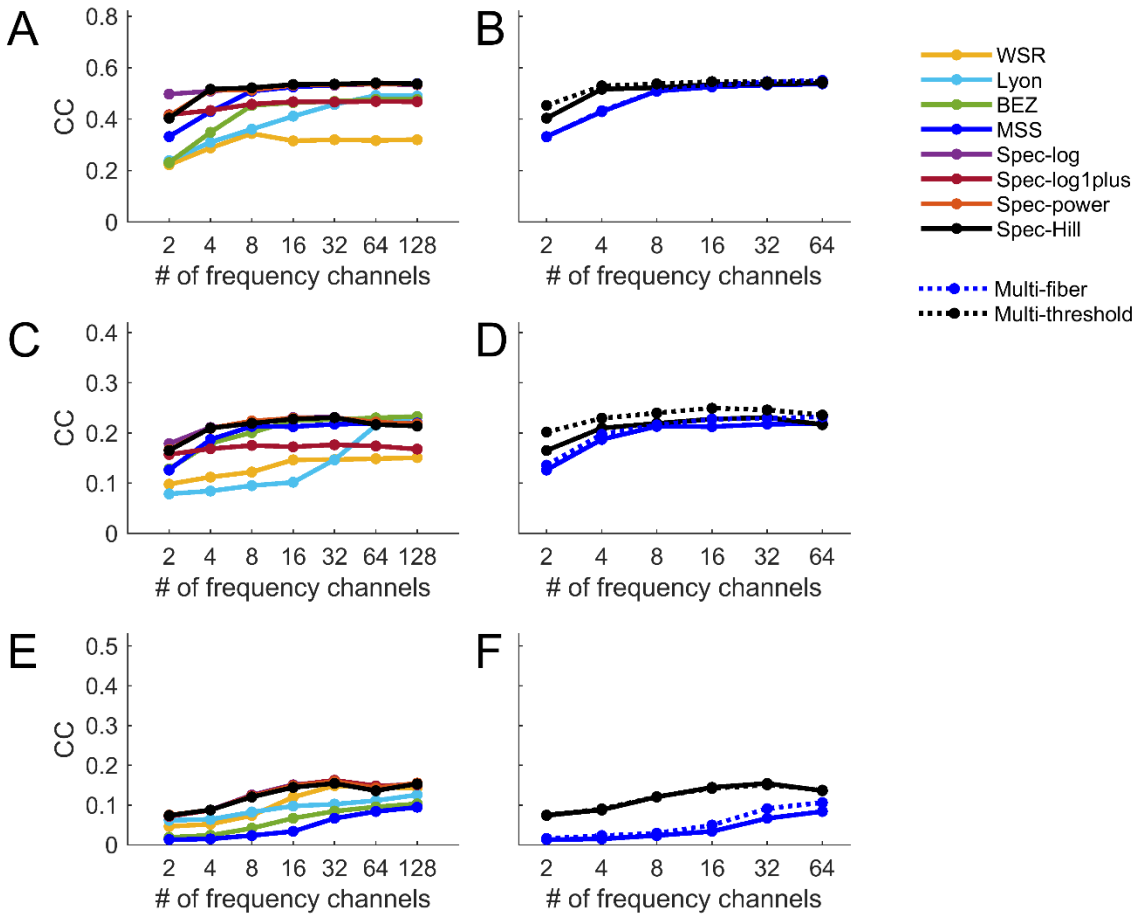

Figure 2.15: Prediction accuracy of models when correlation coefficient (CC) is used as performance measure. *A and B. CC on the natural sound dataset 1. C and D. CC on the natural sound dataset 2. E and F. CC on the DRC dataset.*

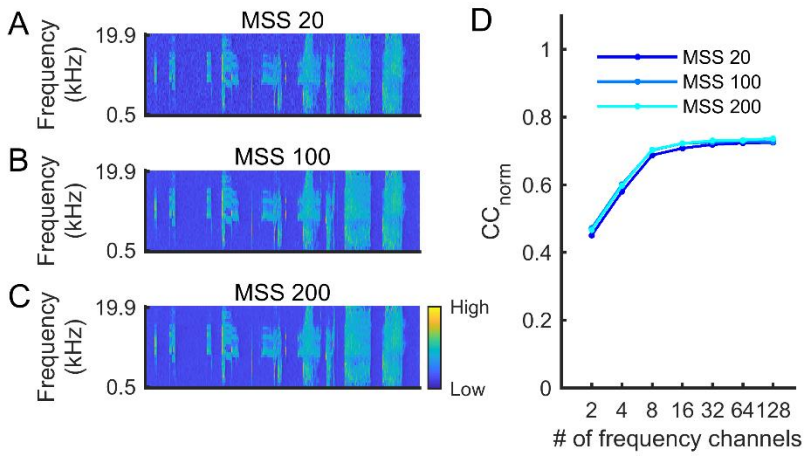

Figure 2.16: Effect of noise on the prediction performance of the MSS model. *A. Cochleagram of a speech sound produced by an MSS model averaged over 20 repeats (MSS 20). B. Cochleagram of the same sound produced by an MSS model averaged over 100 repeats (MSS 100) and C. 200 repeats (MSS 200). D. Comparison of the prediction performance of the MSS 20, MSS 100 and MSS 200 models.*

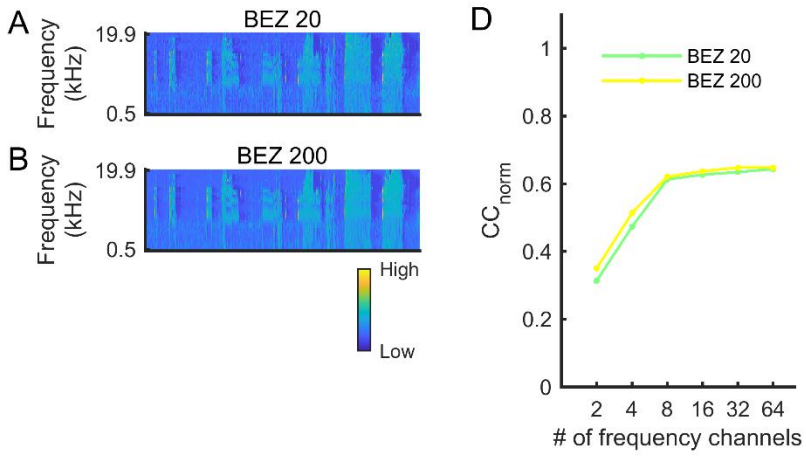

Figure 2.17: Effect of noise on the prediction performance of the BEZ model. *A. Cochleagram of a speech sound produced by a BEZ model averaged over 20 repeats (BEZ 20). B. Cochleagram of the same sound produced by a BEZ model averaged over 200 repeats (BEZ 200) and C. Comparison of the prediction performance of the MSS 20 and the BEZ 200 model.*

The BEZ model takes much longer to run than the other model. For the sake of running time, we ran the BEZ 200 model only up to 64 frequency channels.

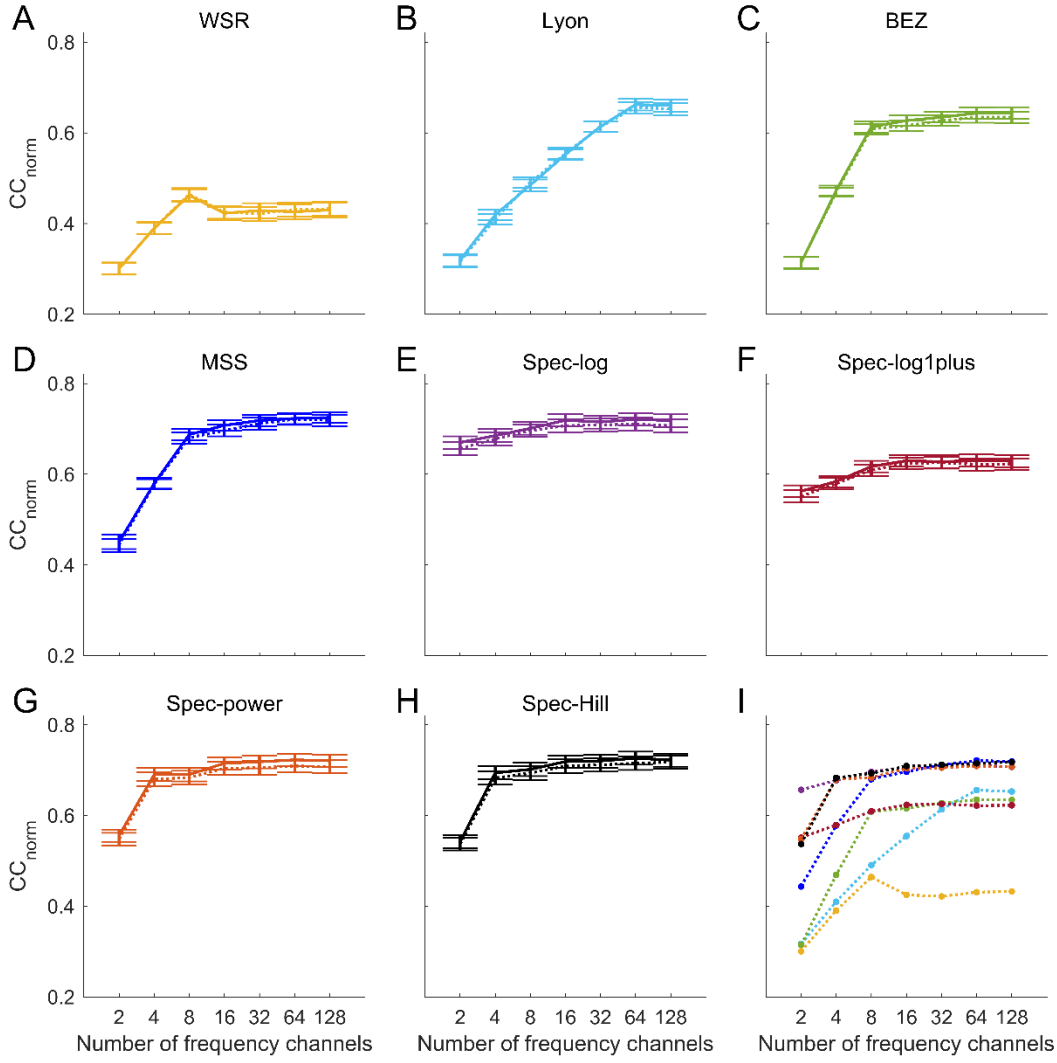

Figure 2.18: Average CC_{norm} of different cochlear models when onset response is included for prediction. *A-H. Solid lines are without onset response (without 0-800 ms), as in Figure 3. Dotted lines are with the onset response included. Error bars on solid and dotted lines are the standard error. I. The CC_{norm} of all the different cochlear models with onset response included.*

We also explored what non-linear aspects of the spectrogram cochlear model and LN model combination are important for good prediction of cortical neural responses. Both the cochlear model and the LN encoding model include non-linearities. To examine how these two non-linearities interact, we constructed a spectrogram cochlear model without any compressive cochlear non-linearity (spec-lin) and compared its performance with the other spectrogram models that included a compressive cochlear non-linearity, in the presence or absence of the LN model output non-linearity. Although there is some variation across nonlinearities and stimulus sets, in general we found that the compressive cochlear non-linearity and the output non-linearity contributed to prediction partially independently. When a model had both nonlinearities together, it typically predicted better than just one nonlinearity on its own, but a compressive cochlear non-linearity tended to contribute more than the output nonlinearity (Fig. 2.19).

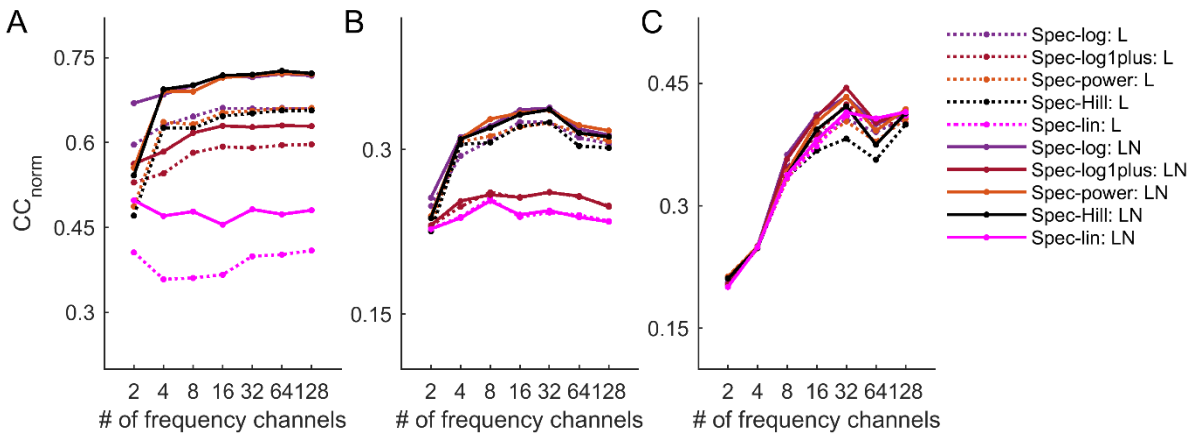

Figure 2.19: Interaction between compressive cochlear non-linearities and LN-model output non-linearities. *Average CC_{norm} of spectrogram-based models (spec-lin is the spectrogram-based model without any compressive cochlear non-linearity) for an LN encoding*

model with (LN) and without (L) the output non-linearity. A. CC_{norm} on the natural sound dataset 1. B. CC_{norm} on the natural sound dataset 2. C. CC_{norm} on the DRC dataset.

Finally, we have extended our analysis beyond the average CC_{norm} for a whole dataset, by exploring how the predictive capacity of the model fits depend on different features of the neurons or stimuli. To display the relative performance of individual neurons, scatter plots of the CC_{norm} of every neuron for each model, plotted against the MSS model, are given for all three datasets in Figure 2.20. Each plot also shows the p value of paired t-test between the CC_{norm} values for two models. Neurons vary in their noise ratio and for the natural sounds, CC_{norm} showed small dependence on noise ratio (Pearson correlation coefficient with significant P-value is negative for most models across different datasets)) - with model performance getting worse for higher NR (Fig. 2.21). We also examined how CC_{norm} depended on the neuron's best frequency and IE score (17, 19). No strong relationships were apparent (Figs. 2.22 and 2.23). When we examined the dependence of CC_{norm} on latency, it did appear that longer latency neurons had lower CC_{norm} values (Fig. 2.24). This is consistent with them perhaps having additional nonlinearities due to receiving stronger inputs from higher cortical areas, as suggested by their long latency (Fig. 2.24). We also explored how well neural responses were predicted for the four stimulus types in the test set of natural sound dataset 1: ferret vocalizations, other animal sounds, speech, and environmental sound. The spec-log/power/Hill models and the MSS model remained consistently among the top performing models for each stimulus type, indicating that the robustness of the spec-log/power/Hill models is not driven by a subset of stimuli (Fig. 2.25). Finally, to explore which aspects of the neural response are better predicted by the different cochlear models, we examined how the mean squared error (MSE) of the model's estimate of the cortical response depended on the recorded spike probability (Fig. 2.26). We found that models that performed well

tended to have lower MSE during the high spike probability times as compared to models that performed less well, suggesting that predicting peaks in high spike activity accurately is a factor in determining model performance.

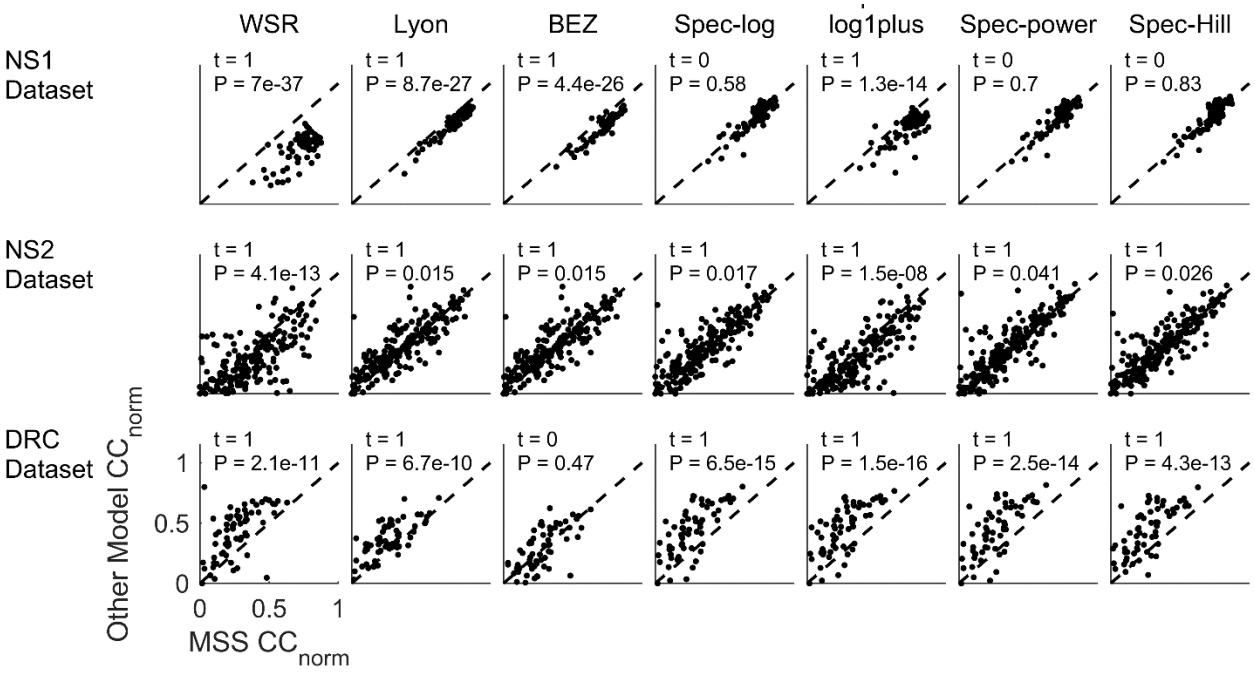

Figure 2.20: Neuron by neuron comparison of model performance. *Each model with its optimum number of frequency channels was chosen for this analysis. Each panel shows the scatterplot of the CC_{norm} of each neuron for a model (model name on top) versus the MSS model. Each panel also shows the result of a pairwise t-test between the CC_{norm} values of two models. Rows are different datasets. Each dot is a neuron.*

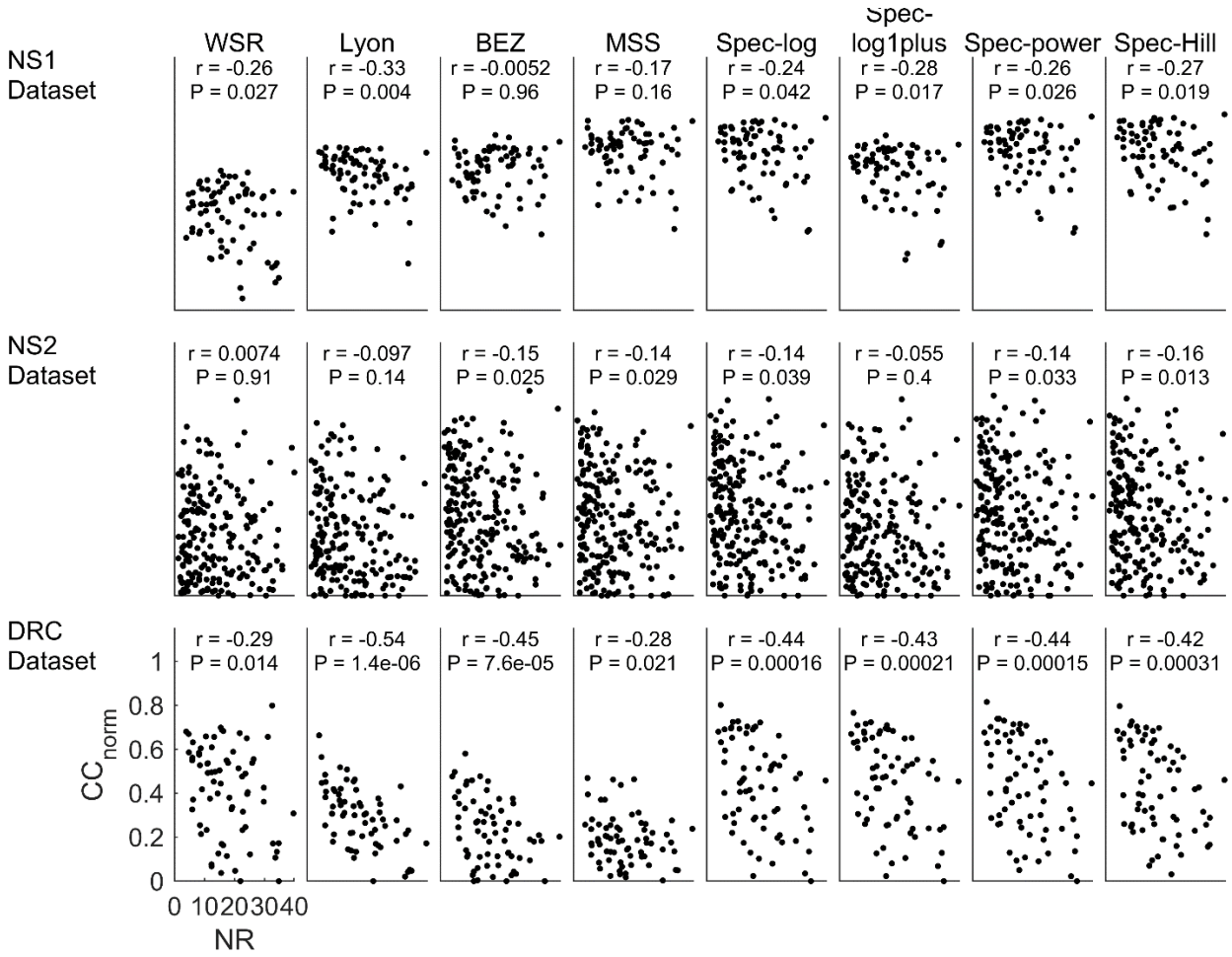

Figure 2.21: CC_{norm} as a function of noise ratio, for the three datasets and all the models. *Each panel shows the scatterplot of the CC_{norm} of each neuron for a model (model name on top) versus the noise ratio (NR) of each neuron. Each panel also shows the Pearson Correlation Coefficient, r, between the two variables and its P-value. Each model with 32 frequency channels was chosen for this analysis. Rows are different datasets. Each dot is a neuron.*

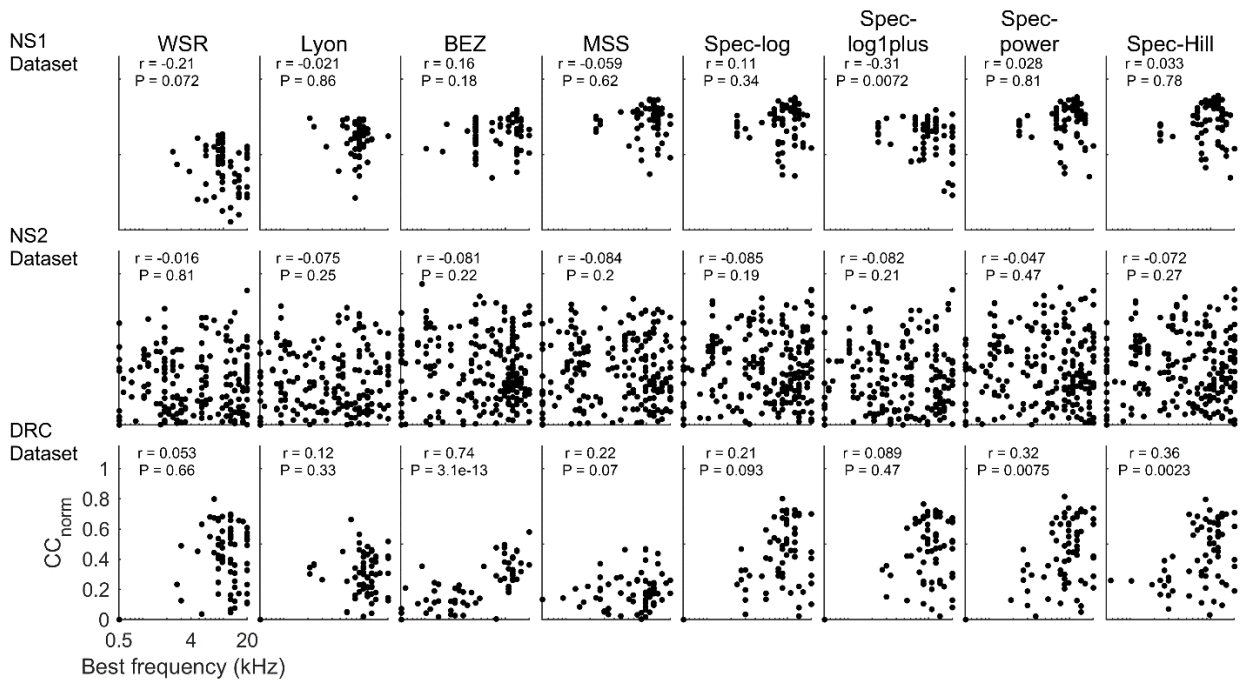

Figure 2.22: CC_{norm} as a function of best frequency of the STRF, for the three datasets, and all the models. *Each panel shows the scatterplot of the CC_{norm} of each neuron for a model (model name on top) versus the best frequency of each neuron estimated from the STRF. Each panel also shows the Pearson Correlation Coefficient, r, between the two variables and its P-value. Best frequency was measured as the frequency field that produced the maximum absolute value in the STRF of a neuron. Each model with 32 frequency channels was chosen for this analysis. Rows are different datasets. Each dot is a neuron.*

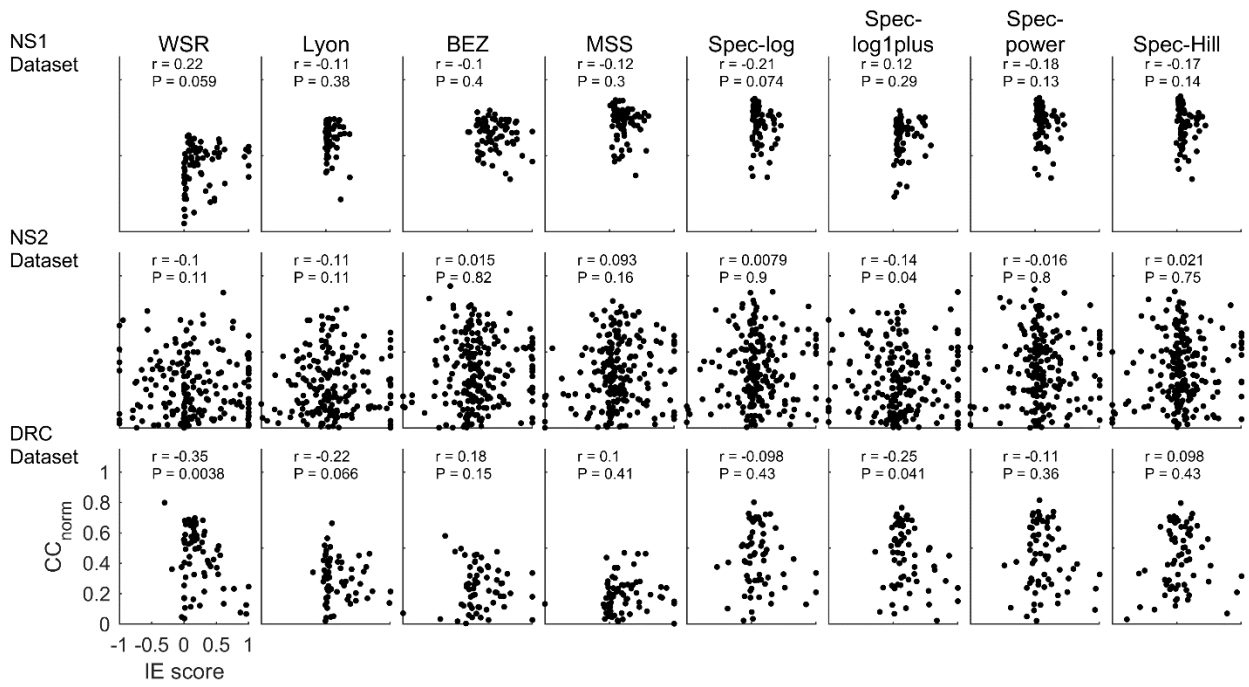

Figure 2.23: CC_{norm} as a function of IE score of the STRF for the three datasets, and all the models. *IE scores measures the difference in strength of excitatory (positive) weights and inhibitory (negative) weights in an STRF (Harper et al., 2016; Rahman et al., 2019). A value of 1 indicates all excitatory weights, a value of -1 indicates all inhibitory weights. Each panel shows the scatterplot of the CC_{norm} of each neuron for a model (model name on top) versus the BF of each neuron estimated from the STRF. Each panel also shows the Pearson Correlation Coefficient, r, between the two variables and its P-value. Each model with 32 frequency channels was chosen for this analysis. Rows are different datasets. Each dot is a neuron.*

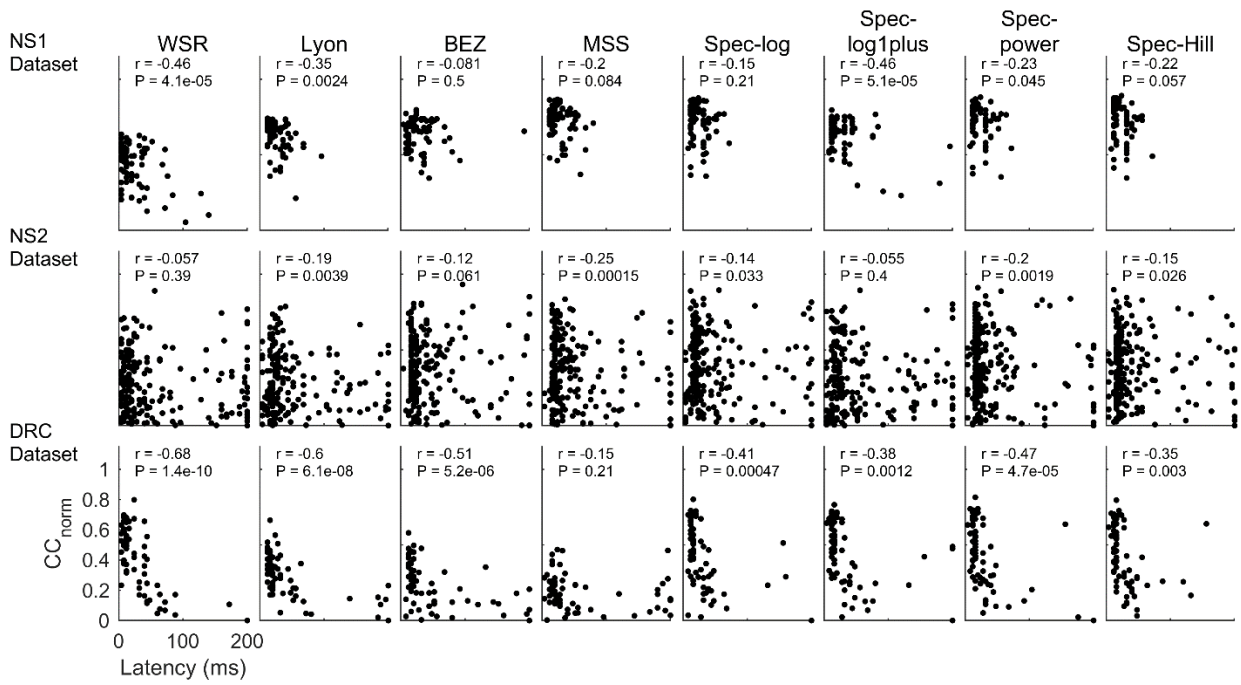

Figure 2.24: CC_{norm} as a function of the latency of the STRF, and all the models. *Each panel shows the scatterplot of the CC_{norm} of each neuron for a model (model name on top) versus the latency of each neuron estimated from the STRF. Each panel also shows the Pearson Correlation Coefficient, r, between the two variables and its P-value. Latency was measured as the time delay that had maximum absolute value in the STRF of a neuron. Each model with 32 frequency channels was chosen for this analysis. Rows are different datasets. Each dot is a neuron.*

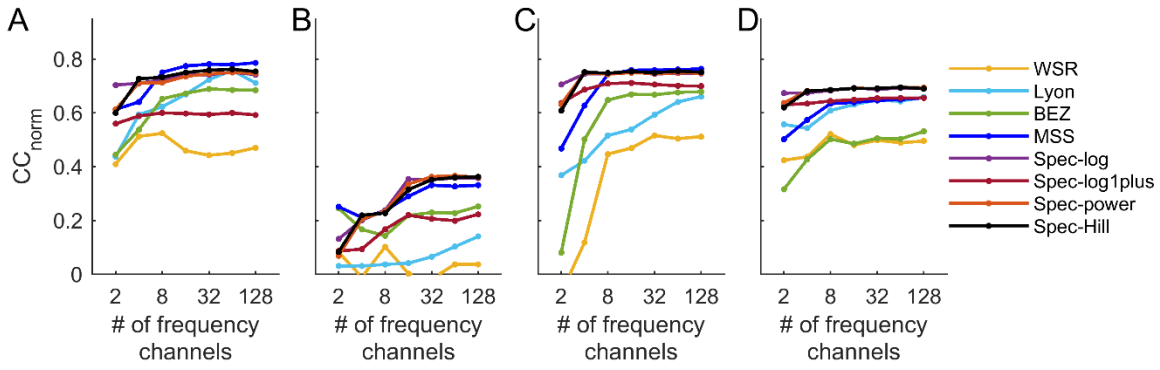

Figure 2.25: CC_{norm} as a function of number of frequency channels, for each cochlear model, for each stimulus type in the test set. *A. Ferret vocalizations. B. Other animal sounds. C. Speech. D. Environmental sounds.*

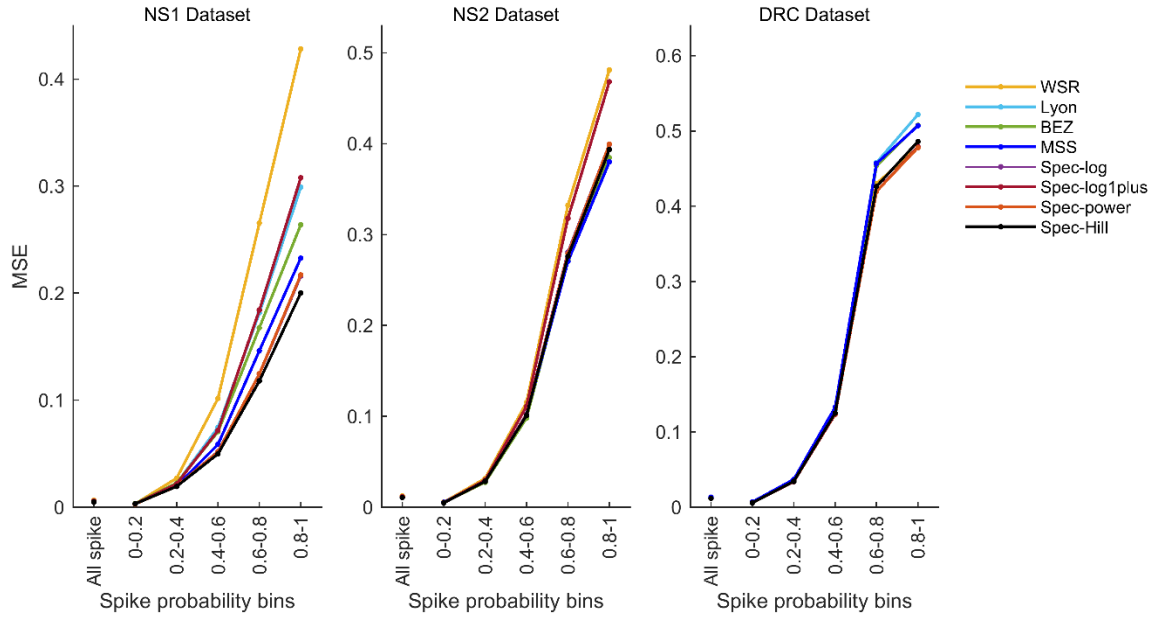

Figure 2.26: Mean squared error (MSE) of model predictions as a function of recorded spike probability. *Each panel is a different dataset, each color is a different cochlear model. For each dataset and model, the MSE between the model prediction and the real PSTH was measured, as function of the spike probability in the PSTH. "All spike" is the MSE over all spike probabilities.*

2.3 Discussion

In this study, we aimed to uncover the computational transformation of the auditory signal from the ear to the cortex. To do this, we investigated and developed different models of the auditory periphery and assessed their capacity to provide the input to encoding models of the responses of auditory cortical neurons to a range of sounds, including natural sounds. Surprisingly, we found that the only models that consistently predicted the responses of A1 neurons across datasets, stimulus type and brain state were the simple models that were based on little more than a spectrogram and some compression (the models spec-log, spec-power, spec-Hill, which all performed similarly). Likewise, a simple spectrogram-based model that approximated the three fiber types of the auditory nerve (the multi-threshold model) tended to produce better predictions of the neural responses than a complex model with extensive biological detail and the three fiber types. These findings hold when the models were used as input to different encoding models (linear, linear-nonlinear, and network receptive fields (Harper et al., 2016; Rahman et al., 2019)), emphasizing their robustness. These findings suggest that the functional transformation from the ear to the auditory cortex might be simpler than expected and that many of the details of the mechanical and neural properties of the ear, and the tuning properties of the auditory nerve and brainstem, may be of limited relevance to their impact on cortex. This exemplifies the distinctions made by Marr (Marr and Poggio, 1976) between the computational and algorithmic levels of analysis, which in this case may be surprisingly simple, and the implementational level, which is very complex.

The observed changes in encoding model prediction performance with different cochlear models were substantial in size, similar to those resulting from inclusion of features such as gain control (Rabinowitz et al., 2011) and network structure (Harper et al., 2016; Rahman et al., 2019) in encoding models of cortical neurons. The choice of cochlear model is therefore likely to be an

important factor in accounting for differences in prediction performance of similar encoding models reported by different groups (David et al., 2009; Calabrese et al., 2011; Willmore et al., 2016). Furthermore, the addition of multiple fiber-types/thresholds can produce substantial improvements in the performance of the model. Additional improvements occur when multiple fiber-types/thresholds were used in conjunction with network structure to predict the responses of the natural sound dataset 1, with prediction performance reaching a remarkably high CC_{norm} of 0.78 (highest achieved so far for this dataset, compared with (Harper et al., 2016; Willmore et al., 2016; Rahman et al., 2019)) for the multi-threshold model. Nonlinear features often do not improve prediction independently. However, the NRF and multi-threshold nonlinearities appear to be relatively independent, both contributing to prediction when applied together. Similarly, with the single-threshold spectrogram models, the cochlear model compression acts relatively independently of the LN-model output nonlinearity.

We have found that the number of frequency channels in the cochleagram is an important factor in influencing the performance of the cortical encoding models. Spectrogram based models performed relatively well with 32 frequency channels and this performance plateaued (for NS1) or became worse (for NS2 and DRC) for 64 or more frequency channels. The DRC was generated using 34 superimposed pure tones that are logspaced over 5.5 octaves, from 500 Hz to 22627 Hz and so it is unsurprising that the DRC performs worse with 64 or more frequency channels in the cochleagram. However, for the NS2 dataset, the reason for this is not very clear. The Lyon model seems to improve performance even with 128 frequency channels relative to 64 frequency channels, although overall the best performing Lyon model cannot out-perform the best performing spectrogram model. Further investigation with more frequency channels might provide us with clues to how many channels will be necessary to saturate the performance of this model in predicting the responses to the NS2 dataset. This will also require more computation time and

more data. Cochlear implants often use as few as 16-22 channels (Zeng et al., 2008). Although some designs include as many as 120 channels, these often are located in a fewer electrodes (de Melo et al., 2012). While it is a possibility that poor quality of hearing by cochlear implants is due to fewer frequency channels, it could also be the case that it is due to the narrower bandwidth of the frequency channels. Some studies show that adding more temporal cues to cochlear implants increases the quality of speech perception without increasing the frequency channels (Green et al., 2005; Nie et al., 2006). Furthermore, our finding that a model of the auditory periphery with 32 frequency channels is sufficient to produce best prediction could be due to the to the limitation of the LN or NRF model to capture more frequency channels. A model that will capture 100% of the CC_{norm} would be able to provide us with better insights on the requirement of frequency channel numbers in the input to a cortical model.

One reason why the spectrogram models performed both well and consistently across stimulus types may be that the biological models fail to accurately represent the processing that takes place in the mammalian auditory periphery generally or the ferret auditory periphery specifically. The WSR and Lyon models are based on the broad results of animal experiments and designed to match certain human psychophysical percepts. The MSS and BEZ models are derived from detailed guinea pig data and cat data, respectively (Sumner et al., 2003; Meddis, 2006; Meddis et al., 2013; Steadman and Sumner, 2018) and verified with auditory nerve responses to simple stimuli. Although there are relatively few physiological studies of the ferret peripheral auditory system, estimates of cochlear frequency selectivity in ferrets (Sumner and Palmer, 2012; Sumner et al., 2018) are comparable to those made for other mammalian species, notably guinea pigs and cats. Recent work suggests that frequency selectivity of ferrets more closely resembles that of guinea pigs than cats (Sumner and Palmer, 2012). Hence, one might expect the MSS model to perform best, but this is only the case for natural sound dataset 1, not natural sound dataset 2

or the DRC dataset. The biological models will have had their many complexities adjusted to account for physiological or psychophysical phenomena based on simple stimuli. Hence, it may be that each model only predicts cortical responses well for stimuli containing features around which the model was constructed. Our simple spectrogram models, in contrast, are not specialized for capturing particular stimulus features. This perhaps renders the simple spectrogram models more representative of the transformation performed by the periphery across a broad span of sounds, including natural sounds.

We observe that the performance of a cochlear model varies between two natural sound datasets used in our study. There are several factors that could have caused this difference of model performance across datasets including the presence or absence of anesthesia, difference of loudness of the sound stimuli (NS1 average sound intensity 80 dB SPL, while NS2 average sound intensity 54 dB), and the amount of training data (NS1 had 16 5 s clips in the training set, while NS2 had 14 4 s clips). One or several of these factors could be behind the variability of prediction performance of a model. Each model's performance is generally worse for NS2 dataset than it is for the NS1 data perhaps due to increased non-linearity in the awake nervous system (Massoudi et al., 2015).

Another possible reason why the spectrogram models perform better than the biological models has to do with the fact that the input to the cortex does not come directly from the auditory nerve. Considerable processing takes place along the auditory pathway, with neurons at each stage being increasingly low-pass to fine-structure (Wallace et al., 2000, 2002, 2011) and amplitude modulation (Joris et al., 2004) and becoming invariant to various features of the stimulus such as acoustic background noise (Rabinowitz et al., 2013). This results in cortical responses being

relatively insensitive to the fine temporal structure and fast amplitude modulation of sounds that are precisely encoded by auditory nerve fibers (Wallace et al., 2000, 2002, 2011; Joris et al., 2004). Furthermore, stimulus features can be transformed nonlinearly into quite different representations at higher levels of the auditory pathway (Joris et al., 2004). This processing may explain why the resulting transformation from ear to cortex is better captured by our simpler spectrogram models than by the cochlear models with more extensive detail of auditory nerve responses.

It is important to consider a few caveats for our results. The dependence of model predictions on stimulus set (David et al., 2009; Theunissen and Elie, 2014; Willmore et al., 2016) and encoding model (Thorson et al., 2015; Meyer et al., 2017) is well recognized. We have shown the robustness of our simple cochlear models over three datasets that differ in stimulus type (natural sounds and DRCs) and brain state (awake and anesthetized), and using three different encoding models (L, LN, and NRF). However, other factors, such as spatial hearing cues were not included, and reverberation, background noise, and sound mixtures were only present to a limited extent. Furthermore, a proportion of the stimulus-dependent neural response in A1 could not be explained by any of the models. This is particularly the case for natural sound dataset 2, perhaps due to increased non-linearity in the awake nervous system (Massoudi et al., 2015), and for the dynamic random chords, perhaps due to more spectral detail. This all implies that while our simple models capture much of the transformation from ear to cortex, a more accurate approximation of the transformation, and one that applies more widely, may be more complex. Investigating which aspects of auditory processing at subcortical levels of the auditory pathway are relevant to models of cortical neurons can be determined empirically by similar methods. For example, our results suggest that the division of the auditory signal among the three physiologically distinct categories of auditory nerve fiber is an important detail for the ear-to-cortex transformation.

Our study is the first extensive comparison focused on the capacity of different peripheral models to capture cortical neural response in mammals. In a pioneering earlier study in birds, Gill et al. (Gill et al., 2006) examined how well different cochlear models predicted neural responses to conspecific birdsong and modulation-limited noise in the avian midbrain and the primary and secondary auditory forebrain, which are considered to be the avian homologue of mammalian A1 (Calabrese and Woolley, 2015). They found that time-frequency scale and whether logarithmic or linear frequency spacing of filters was used had limited impact, and that the optimal values in the models were stimulus dependent. The time-frequency scale relates to the number of frequency channels (although it also relates to time resolution, complicating matters). Over an equivalent range of frequency channels (about 20-120), we similarly found that the number of channels often had limited impact on prediction and that the optimal number depended on the stimulus and model. However, below ~20 channels we generally found more channels to be better. As in our study, Gill et al. (Gill et al., 2006) also found that sub-logarithmic compression, linear and log(1+(.)) in our case, linear and power law in theirs, fitted neural responses worse than logarithmic. Similarly, in an investigation of linear encoding models of mammalian A1 neurons (Thorson et al., 2015), which also explored channel number, the peripheral model selected used near-logarithmic compression and 18 channels. This is consistent with our study and that of Gill et al. (Gill et al., 2006) in showing that a model incorporating logarithmic compression and at least ~20 frequency channels predicts cortical responses well.

Finally, in contrast to our study, Gill et al. (Gill et al., 2006) found that Lyon's model with adaptive gain-control provided the best fit to neural responses for both their stimulus types. Reasons for this difference could be the species or stimulus sets used, or details of our spectrogram models such as the use of triangular filters. Gain control and other adaptive phenomena are ubiquitous

features of mammalian auditory processing (Fritz et al., 2003; Dean et al., 2005, 2008; Rabinowitz et al., 2011; Maier et al., 2012; Robinson et al., 2016), so the Lyon model's underlying transmission-line cochlear model or the parameters of its adaptive gain control may simply not match well with the adaptation and other features exhibited by the ferret auditory system. An appropriate cochlear model with species-specific gain adaptation may further improve on our results, particularly as adding sound-level adaptation to our spectrogram-based models improves prediction of ferret cortical responses (Willmore et al., 2016).

Relevant to our multi-threshold model is a model of A1 responses that also uses multiple level-dependent input nonlinearities (Ahrens et al., 2008). This input nonlinearity model transformed the sound levels in each frequency band using a set of fixed basis functions. The basis functions are not biologically inspired in contrast to our multi-threshold model, and were applied to DRCs with ten discrete sound levels, rather than natural sounds and continuous-valued DRCs. This model predicted rodent A1 responses to DRCs better than an STRF model, but it was not compared with an LN model and the interaction of model's input nonlinearity with an output nonlinearity was not explored. Finally, single hidden layer artificial neural networks have been applied to predict ferret A1 responses (Harper et al., 2016). This network model also has features resembling multiple input nonlinearities (the hidden unit nonlinearities), albeit with a linear transform applied first and this was shown to predict better than an LN model. However, as we have seen in Figure 5, appending this network model to our multi-threshold model even further improves prediction of A1 responses to natural sounds, indicating that the multiple thresholds and the hidden unit nonlinearities capture different nonlinear aspects of the ear to cortex signal transformation.

It is interesting to speculate on the perceptual, behavioral and clinical implications of our findings. Algorithms for automatically estimating speech quality (and other speech characteristics) are useful for assessing hearing aids and other auditory prosthetics and can also indicate what sound transformations guide perception and action. Wirtzfeld et al. (Wirtzfeld et al., 2017) found that a complex biologically-detailed cochlear model (Zilany et al., 2009, 2014) did not outperform a simpler model (Kates and Arehart, 2014) in predicting human estimates of speech quality in noise, although the relative effectiveness of different models in assessing speaker identity is affected by the presence of different types of background noise (Islam et al., 2016). Similarly, a simple cochlear model is sufficient to reproduce the task-dependent STRF plasticity that characterizes auditory cortex (Chambers et al., 2019). If cortical activity reflects perception and guides behavior, these studies are consistent with our finding that a simple transformation predicts cortical responses well and also suggests possible value in using our simple models in speech assessment and recognition algorithms. An alternative way to investigate what aspects of cochlear models are important for perception would be to synthesize sound textures (McDermott and Simoncelli, 2011) using different cochlear models, and quantitatively assess human judgements of their quality. Our findings also have implications for cortical implants (Lim et al., 2009), by suggesting simple signal processing strategies to mimic the impact of the auditory periphery on the stimulated neurons.

In summary, although extensive processing takes place in the cochlea and the central auditory pathway (Schnupp et al., 2011), our results suggest that the cortex receives a relatively simple functional transformation of sound inputs. Although the complex properties of the auditory periphery are important for auditory processing, one possibility is that many such properties could be present in the system due to the physical constraints of cellular and physiological machinery. Many of the complex properties of peripheral auditory processing appear to have limited impact

on cortical responses, and much of that processing is captured by a simple spectral decomposition of the input. Explaining the remaining aspects of how cortical neurons respond to natural sounds will likely require additional complexities to those found in the models we examined, which can be revealed using empirical computational methods such as those adopted here. It is likely that similar principles apply to other sensory systems.

2.4 Methods

Neural responses to sound stimuli were recorded from ferret primary auditory cortex. The sounds were put through a cochlear model which then provided input to an encoding model. The parameters of the encoding model were optimized to estimate the time-course of the neural responses to the sounds, and the model was then tested on how well it could predict responses to sounds that were not used for the optimization. The average capacity of different cochlea models to predict the neural responses via encoding models was examined, to determine which cochlear model best captured the impact of cochlear processing on the neural responses.

2.4.1 The dataset: stimuli and neural response

Three datasets were used in this study: natural sound dataset 1 (NS1), natural sound dataset 2 (NS2), and the dynamic random chord (DRC) dataset. Each dataset consisted of a number of sound clips, and the neural responses to these clips.

Natural sound dataset 1 (NS1)

This dataset was gathered by the authors and used in previous studies, in some cases as the sole dataset in a study (Harper et al., 2016; Rahman et al., 2019) and others as part of the dataset [experiment 1 in (Schoppe et al., 2016); part of the “big nat” dataset in (Willmore et al., 2016)]. The dataset is publicly available at https://osf.io/ayw2p/. All data were obtained from experiments performed under license from the UK Home Office and approved by the University of Oxford Committee on Animal Care and Ethical Review. They were obtained using 16- or 32-channel silicon probe electrodes (Neuronexus Technologies) from anesthetized ferrets (ketamine, 5 mg/kg/h; medetomidine, 0.022 mg/kg/h). *In vivo* extracellular recordings were made from the primary auditory cortical areas (primary auditory cortex, A1 and anterior auditory field, AAF, which

we collectively refer to as A1) in response to natural stimuli. The sounds were presented binaurally through earphones (Panasonic RPHV297) attached to otoscope speculae inserted into the ear canals, using a sampling rate of 48828.125 Hz. The sound stimuli consisted of 20 clips of natural sound (speech, ferret vocalizations, other animal vocalizations, and environmental sounds), each 5 s in duration. The clips were played in random order. Each clip was repeated 20 times. The RMS sound intensity over the full 5 s duration of each clip ranged from 75 to 82 dB SPL. The RMS sound intensity within 4 ms time bins varied from 29 dB SPL to 105 dB SPL over the whole dataset. In total, the dataset included 549 single and multi-unit recordings from six adult pigmented ferrets (five female and one male), among which 284 were single units. Seventy-three of the single units had noise ratios (Linden et al., 2003; Rabinowitz et al., 2011) to the stimulus set of < 40, and these were the units used in this study.

Natural sound dataset 2 (NS2)

The second natural sound dataset has been published by other authors (Espejo et al., 2019) and is publicly available (https://zenodo.org/record/3445557#.XyNEy-d7lEY). A description of the experimental setting and stimuli can be found in (Espejo et al., 2019). Briefly, extracellular neurophysiological recordings were obtained from (A1 and AAF) of six awake ferrets using 1–4 independently positioned tungsten microelectrodes while the animals were passively listening. Sound was delivered through free-field speakers (Manger W05) positioned at ±30° azimuth and 80 cm distant from the animal, at a sampling rate of 44.1 kHz. The sound stimuli consisted of diverse clips of natural sounds. Most clips were presented only once, but some were repeated a number of times, which varied over experiments. In a subset of experiments, there were 18 natural sound stimuli clips (each 4 s in duration) that were repeated 10 (or sometimes 20) times. We selected this subset of experiments and used the neural responses to these 18 natural sounds for our analysis, as a sufficient of repeats is required for the calculation of noise ratio and the

PSTH, and a sufficient number of sound clips is needed for model fitting. This set of 18 natural sound clips included human speech, ferret and other species' vocalizations, natural environmental sounds, and sounds from the animals' laboratory environment. The RMS sound intensity of the full 4 s duration of each clip was 54 dB SPL. The RMS sound intensity within 4 ms time bins varied from -16 dB SPL to 76 dB SPL over the whole dataset. The neural dataset for these 18 stimuli consisted of 336 single units. We used the 235 of these units, selected because they had a noise ratio <40 for the 18 natural sound stimuli.

The DRC dataset

The dynamic random chord (DRC) dataset comes from the same anesthetized ferret experiments as the natural sound dataset 1. The DRC dataset has not previously been analyzed in publications, but is publicly available at https://osf.io/ayw2p/. The DRC stimulus clips were randomly interleaved between the clips in natural sound dataset 1, hence the units in the DRC dataset are the same 73 single units as in that natural sound dataset. The sound stimuli in the DRC dataset consisted of 12 sound clips each of 5 s duration, and each clip was presented 5 times. Each clip was of dynamic random chords (deCharms et al., 1998; Schnupp et al., 2001; Linden et al., 2003; Willmore et al., 2016), which consisted of a sequence of complex tones, each tone being 25 ms long, presented with no gap between the tones. Each complex tone was composed of 34 superposed pure tones, each of whose levels were independently and randomly chosen. The frequencies of the 34 pure tones were log spaced over 5.5 octaves, from 500 Hz to 22627 Hz, with 1/6 octave spacing. The tone levels were picked from a uniform distribution of spanning either 10, 20, 30 and 40 dB, with a mean of 44 dB SPL. The RMS intensity over the full 5 s duration of each clip ranged from 79 to 88 dB SPL. The RMS sound intensity within 4 ms time bins clips varied from 71 dB SPL to 93 dB SPL over the whole dataset.

2.4.2 Cochlear models

Figure 2.1 illustrates the similarities and differences between of the cochlear models that we used. Below we describe these models.

WSR model: This model was originally proposed by Wang and Shamma and was later described by Powen Ru (Matlab codes that we are using were adapted from https://github.com/tel/NSLtools) (Wang and Shamma, 1994, 1995; Ru, 2001; Chi et al., 2005). We refer to it as the WSR (Wang Shamma Ru) model, after the names of the developers. It makes use of a series of filters, whose center frequencies were spaced logarithmically, followed by a sigmoid compression of the filter outputs. The sigmoid compressed outputs were then passed through a lowpass filtering stage, a lateral inhibition stage (between neighboring frequency channels), and a temporal integration stage. The source of these parameters is not strictly physiological, rather, they are abstracted from animal experiments to match perceptual processing. The filterbank of this model comprised a set of 129 given frequency channels. To make this model comparable to other models, appropriately spaced and sized subsets of the frequency channels were selected.

Lyon model: This model uses a cascade of filters with half-wave rectification and adaptive gain control whose center frequencies are spaced according to a logarithmic scale, except for low frequencies, where the spacing is linear, to simulate the behavior of the cochlea (Lyon, 1982, 2011). The parameters of the filters have been largely obtained from human psychophysics experiments. We adapted this model from: https://github.com/google/carfac. This implementation has several free parameters, allowing for flexible choice of spacing and bandwidth of each

frequency channel in the filterbank – we set these parameters to values that would generate the desired number of frequency channels between 0.5 Hz and 20 kHz.

BEZ model: This is a phenomenological model of cat auditory nerve fibers (Zilany et al., 2009, 2014; Bruce et al., 2018), which we refer to as the BEZ (Bruce Erfani Zilani) model [the model described in (Bruce et al., 2018)], named after its developers (we adapted this model from: https://www.urmc.rochester.edu/MediaLibraries/URMCMedia/labs/carney-lab/codes/UR_EAR_v2_1.zip). This model has various processing stages to match the processing stages in the cochlea, the synapses with the auditory nerve and the spiking properties of the auditory nerve. Sound input is first processed through a middle ear filter, followed by a bank of filters with various complexities and non-linearities to account for inner hair cell and outer hair cell properties, the activity of the hair cells is then processed through a nonlinear model of synaptic neurotransmitter release, uptake of neurotransmitter and spiking of the auditory nerve. Because the spiking is a stochastic process, for each center frequency, each of three auditory nerve fiber types was allowed to spike many times (200 for plotting cochleagrams, and 20 for predicting neural responses to reduce running time) and the average was taken over these trials. After that, we took an average over different fiber types weighted by the ratio of each fiber type in the nerve fiber population.

MSS model and multi-fiber model: Different components of this model have been incorporated and refined over the years; we call it the MSS (Meddis Sumner Steadman) model after the name of its developers (we adapted this model from: https://zenodo.org/record/1345757#.XW4XtXt7IPY). The MSS model is a biologically-detailed model of inner hair cells in the cochlea and the auditory nerve (Meddis et al., 2001, 2013; Sumner et al., 2002, 2003; Meddis, 2006; Steadman and

Sumner, 2018), the parameters of which are based on measurements from the guinea pig (Meddis et al., 2001; Sumner et al., 2002). The MSS model includes several stages of processing. The first stage – an outer and middle ear (OME) model – uses a series of linear filter approximation to mimic stapes motion at the oval window. The second stage, the frequency decomposition stage, uses a dual resonance non-linear (DRNL) filter architecture to model the properties of the basilar membrane. A DRNL filter consists of two parallel pathways comprising a series of bandpass and lowpass filters, one with a nonlinear compression and the other without a nonlinear compression. The outputs of each parallel pathway are then combined to find the velocity of the basilar membrane displacement. Finally, transduction by the inner hair cells is modelled using a differential equation for their membrane potential, the synapse is modelled using probabilistic release of neurotransmitter and the activity of the auditory nerve (AN) fibers involves refractoriness, with the spiking of the AN fibers providing the output (Fig. 2.1A-C). We set the center frequency of the non-linear filterbank to be log spaced between 508 Hz and 19,912 Hz. Like the BEZ model, for each center frequency each of three auditory nerve fiber types was allowed to spike many times (200 for plotting cochleagrams, and 20 for predicting neural responses to reduce running time) and the average was taken over these trials. For the main MSS model, the output was averaged over the three fiber types, whereas for the multi-fiber MSS model this was not done. Finally, the output was down-sampled into 4-ms time-bins to find a windowed time-frequency representation.

Spec-log model: A spectrogram was produced from the sound waveform by taking the amplitude spectrum using 8-ms Hanning windows, overlapping by 4 ms. The amplitude of adjacent frequency channels was summed using overlapping triangular windows (using code adapted from melbank.m, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html), with log-spaced center frequencies ranging from 508 Hz to 19,912 Hz. The amplitude in each time-frequency bin was

converted to log values (using $20 * \log(\cdot)$) and any value below a threshold was set to that threshold (threshold is -40 for NS1 and DRC, -50 for NS2, set by cross-validation).

Spec-log1plus model: This model is the same as the spec-log model but uses log(1+(·)) as the compression function instead of log(·).

Spec-power model: This model too is the same as the spec-log model, but takes the square of the spectrogram, resulting in a power spectrogram instead of an amplitude spectrogram. This model uses a logarithmic compression function (10 * log(·)).

Spec-Hill model: This model is an extension of the spec-power model where the output of spec-power model is further compressed using a non-linear Hill function (Sachs and Abbas, 1978). Before the spec-power output was put into the Hill function, the magnitude of the spec-power model's threshold was added to this output to ensure that it was non-negative. The Hill function is given by,

$$f(\cdot) = \frac{(c*(\cdot))^{1.77}}{\theta+(c*(\cdot))^{1.77}} \ldots (1)$$

Here, c is the scaling factor, which was set to 0.01 and θ is the saturation parameter of the Hill function, which was set to 0.16. These values of c and θ were chosen by cross-validation on the validation set for natural sound dataset 1, and then left unchanged for the other datasets (see below).

Multi-threshold model: The multi-threshold version of the spec-Hill model was constructed by taking the output of log compression in a spec-power model but applying three different thresholds of -40, -30 and -15 and thresholding it using values of 0 10 and 25 dB (Colburn et al., 2003), and then compressing it with a Hill function (Equation 1) with different saturation levels (saturation was achieved by using Equation 1 with various θ values: $3x10^{-5}$, $5x10^{-5}$, $8x10^{-5}$ and c = $1x10^{-4}$). The threshold, θ and, c values were chosen to be comparable to biological measurements, and then θ and c were slightly adjusted by cross-validation for natural sound dataset 1. These values were then left unchanged for the other datasets (see below).

2.4.3 The PSTH and normalized cochleagram

For each neural unit in a dataset and for each sound stimulus clip n, a peri-stimulus time histogram (PSTH) was made. To construct the PSTH the number of spikes was counted in consecutive 4-ms time bins over the course of the clip and averaged over stimulus repeats. The PSTH to stimulus n is henceforth denoted as $y_n[\mathrm{t}]$, where t indicates the time bin and goes from t = 1 to T. During the beginning of a clip presentation the neuron is adapting from silence, which can be a challenge to model. It is therefore common practice in fitting STRF and LN models to not use this period of sound (Depireux et al., 2001; Harper et al., 2016; Willmore et al., 2016). Hence, the first 800 ms of the neural response for each clip was clipped from the data for natural sound dataset 1 (although including the first 800ms made very little difference to the results, see Figure S12). This clipping was not done for natural sound dataset 2 or the DRC dataset as these clips were shorter than those of natural sound dataset 1 and this clipping would have shortened them too much.

The output of each cochlear model is called a cochleagram: the frequency-decomposed transformation of sound over specific time windows (4 ms in our case). To provide input for an encoding model of an auditory cortical neuron, the cochleagrams were normalized to zero mean and unit variance and for each snippet *n*, for every time *t*, a time-lagged matrix was extracted from the cochleagram. This was done according to the equation, $z_{fqn}[t] = k_{fn}[t - q + 1]$, where *k* is the input vector, *f* is the frequency channel, which goes from *f* = 1 to *F*, *q* is the time lag, which goes from *q* = 1 to the maximum time lag Q, and $k_f[t]$ is the cochleagram. The resulting tensor $z_{fqn}[t]$ is the input to the linear nonlinear (LN) stage of the encoding scheme. Here, *Q* = 50, which is 200 ms. For simplicity of notation, the subscript n that indicates clip number will be dropped unless otherwise stated, rendering the cochleagram tensor as $z_{fq}[t]$ and the PSTH as $y[t]$.

2.4.4 The encoding models

The spectro-temporal receptive field and the linear-nonlinear model

The linear-nonlinear (LN) model consists of a linear stage, the spectrotemporal receptive field (STRF), and a non-linear stage. The linear part of the model is:

$$a[t] = \sum_{f,q} w_{fq} z_{fq}[t] + b \quad \ldots (2)$$

where w is a vector of the input weights and *b* is the background activity of the neuron. Both *w* and *b* are free parameters of the model, which were estimated by linear regression of the neuron's firing rate $y[t]$ on the cochleagram tensor $z_{fq}[t]$ using glmnet (Friedman et al., 2010), where $a[t]$ is the linear estimate of $y[t]$. To overcome overfitting, the parameters were regularized using L_1-norm (LASSO) regularization of the weights. A regularization hyperparameter *λ* was used to

control the strength of the regularization, which was chosen by a cross-validation procedure (see below).

The second stage of the model was a logistic activation function (sigmoid nonlinearity), which was fitted after fitting the STRF. This function is given by:

$$\hat{y}[t] = \frac{\rho_1}{1+\exp\left(\frac{-a[t]-\rho_3}{\rho_2}\right)} + \rho_4 \ \ldots (3)$$

The four parameters ρ_i of the function were fitted by minimizing the squared error between the nonlinear estimate of the firing rate $\hat{y}[t]$ and measured firing rate $y[t]$.

The network receptive field model

The network receptive field (NRF) models the neural response using a network, and is the same model as reported in (Rahman et al., 2019) and a slightly modified form of (Harper et al., 2016). This consists of an artificial neural network with a single hidden layer of 20 hidden units (HU) that converge onto an output unit (OU). Each hidden unit of the model is similar to a single LN model, which all then feed into another LN model that predicts the neural response. The activation of the *j*-th HU, $a_j[t]$ is given by,

$$a_j[t] = \sum_{f,q} w_{jfq} z_{fq}[t] + b_j \ \ldots (4)$$

The output of the *j*-th HU, $v_j[t]$ is,

$$v_j[t] = \mathrm{g}\big(a_j[t]\big) \ldots (5)$$

where $\mathrm{g}\big(a_j[t]\big)$ is a sigmoid non-linear activation function, $1/(1+\exp(-a_j[t]))$. The HUs feed these outputs to the OU. The activation function of the OU, $a_o(t)$ is,

$$a_o[t] = \sum_j w_j v_j[t] + b_o \ldots (6)$$

The output $v_o[t]$ of the OU is,

$$v_o[t] = g(a_o[t]) \ldots (7)$$

which is the model's estimate of the neuron's response at time *t*.

The free parameters (the weights w_{jfq} and w_j, and the biases b_j and b_o) were optimized by minimizing the squared error between predicted response, $v_o[t]$, and actual response, *y*[*t*], subject to L_1-regularization of the weights. Thus, the objective function is given by,

$$E = \frac{1}{2NT}\sum_{n,t}\big(v_{o,n}[t] - y_n[t]\big)^2 + \lambda\big(\sum_{j,f,q}|w_{jfq}| + \sum_j |w_j|\big) \ldots (8)$$

Here, *n* is included to indicate sound clip number, but was left out of other equations for simplicity. *N* is the number of clips used in training.

The parameters of the NRF model are fitted by minimizing the objective function with respect to the free parameters using the sum-of-function optimizer algorithm (Sohl-Dickstein et al., 2013). In using this algorithm, we take one clip to be one minibatch. The optimization algorithm requires calculation of the error gradients in respect to each of the parameters. For the NRF model, error gradients are calculated using standard chain rule (backpropagation).

Before training, the weights were initialized by modified Glorot initialization from a uniform distribution ranging from - $\frac{1}{\sqrt{FQ+L}}$ to + $\frac{1}{\sqrt{FQ+L}}$, where FQ is the number of input weights to a HU and L = 1 is the number of output weights from a HU. The biases were initialized similarly (Glorot and Bengio, 2010).

2.4.5 Cross-validation and testing of the encoding models

Each sound dataset was divided into a cross-validation set and a test set. The cross-validation set was used for training the weight matrices of the encoding models and setting their regularization hyperparameter strength by cross-validation. Also, for all cochlear models, biologically-complex and simple, cross-validation methods (using NS1 mostly) were used to explore some of their settings to avoid using any settings that gave poor predictions. The hyperparameters and settings were selected only using the cross-validation set, and then the separate test set was used to assess the prediction capacity of the models.

More specifically, for the natural sound dataset 1, which contained 20 sound clips, 4 were chosen as a test set that was not used during training and cross-validation. The cross-validation set (the remaining 16 clips) was used to fit the models using *k*-fold cross-validation, where $k = 8$. The cross-validation set was randomly divided into a training set of 14 clips and a validation set of 2 clips. The encoding models were trained on the training set for 18 different values of the hyperparameter λ. A log spaced range of lambda values was used, but with a somewhat lower density at the extremes. For the LN model, the exact values of λ used were: 1.00×10^{-1}, 2.00×10^{-2}, 1.17×10^{-2}, 6.84×10^{-3}, 4.00×10^{-3}, 2.34×10^{-3}, 1.37×10^{-3}, 8.00×10^{-4}, 4.68×10^{-4}, 2.74×10^{-4}, 1.60×10^{-4}, 9.36×10^{-5}, 5.41×10^{-5}, 3.20×10^{-5}, 6.40×10^{-6}, 1.28×10^{-6}, 2.56×10^{-7}, and 5.12×10^{-8}. For the NRF model, the exact values of λ used were: 1.00×10^{-3}, 2.00×10^{-4}, 1.17×10^{-4}, 6.84×10^{-5}, 4.00×10^{-5}, 2.34×10^{-5}, 1.37×10^{-5}, 8.00×10^{-6}, 4.68×10^{-6}, 2.74×10^{-6}, 1.60×10^{-6}, 9.36×10^{-7}, 5.41×10^{-7}, 3.20×10^{-7}, 6.40×10^{-8}, 1.28×10^{-8}, 2.56×10^{-9}, and 5.12×10^{-10}. For each LN/NRF model fitted with different λ, neural responses were then predicted for the validation set, and the correlation coefficient between the actual neural responses and the prediction was measured. This process was repeated 8 times for different non-overlapping validation sets, so the whole cross-validation set provided validation. The model was then retrained with the whole cross-validation set using the λ value that provided the highest mean correlation coefficient over all 8 folds. Next, the retrained model was used to predict the neural responses to the test set. All the correlation coefficients and normalized correlation coefficients shown are for this held-out test set, and all the LN model parameters shown are for the retrained model.

For the other datasets, clip numbers in the training, validation and test sets followed similar proportions. For natural sound dataset 2, which contained 18 natural sound clips, 4 were chosen as the test set and 14 as the cross-validation set. These 14 clips were then further divided into a

training set (12 clips) and a validation set (2 clips) – this was repeated 7 times to obtain 7-fold cross-validation with 7 non-overlapping validation sets that spanned the cross-validation set. For the DRC dataset, among the 12 sound clips, 3 were chosen as the test set and 9 as the cross-validation set. These 9 clips were then further divided into a training set (7 clips) and a validation set (2 clips) – this was repeated 5 times to obtain 5-fold cross-validation with 5 mostly non-overlapping validation sets that spanned the cross-validation set. The range of values explored of hyperparameter λ remained unchanged among datasets.

2.5 Supplementary Tables

Table S2.1: Mean ± SE CC_{norm} between cortical responses and predicted responses for the LN model.

	Number of frequency channels						
Models	**2**	**4**	**8**	**16**	**32**	**64**	**128**
Mean CC_{norm} of the models for the natural sound dataset 1							
WSR	0.301 ± 0.013	0.390 ± 0.013	0.462 ± 0.014	0.423 ± 0.015	0.429 ± 0.017	0.426 ± 0.016	0.430 ± 0.016
Lyon	0.320 ± 0.013	0.419 ± 0.012	0.485 ± 0.013	0.552 ± 0.012	0.614 ± 0.012	0.662 ± 0.013	0.660 ± 0.014
BEZ	0.313 ± 0.013	0.473 ± 0.011	0.613 ± 0.013	0.627 ± 0.012	0.635 ± 0.012	0.643 ± 0.012	0.644 ± 0.013
MSS	0.451 ± 0.015	0.580 ± 0.012	0.687 ± 0.012	0.708 ± 0.012	0.719 ± 0.012	0.723 ± 0.012	**0.725 ± 0.012**
Spec-log	0.670 ± 0.014	0.685 ± 0.014	0.701 ± 0.014	0.719 ± 0.014	0.715 ± 0.014	0.721 ± 0.014	0.718 ± 0.014
Spec-log1plus	0.562 ± 0.013	0.584 ± 0.013	0.617 ± 0.012	0.630 ± 0.013	0.627 ± 0.014	0.630 ± 0.013	0.629 ± 0.013
Spec-power	0.556 ± 0.013	0.690 ± 0.014	0.690 ± 0.015	0.715 ± 0.014	0.718 ± 0.014	0.722 ± 0.014	0.720 ± 0.014
Spec-Hill	0.542 ± 0.014	0.694 ± 0.015	0.701 ± 0.015	0.719 ± 0.014	0.720 ± 0.014	**0.726 ± 0.014**	0.722 ± 0.014
Multi-fiber	0.450 ± 0.012	0.584 ± 0.011	0.688 ± 0.011	0.716 ± 0.011	0.736 ± 0.012	**0.742 ± 0.012**	
Multi-threshold	0.610 ± 0.012	0.713 ± 0.012	0.724 ± 0.013	0.736 ± 0.012	0.734 ± 0.012	0.734 ± 0.012	
Mean CC_{norm} of the models for the natural sound dataset 2							
WSR	0.150 ± 0.015	0.172 ± 0.015	0.183 ± 0.016	0.219 ± 0.015	0.219 ± 0.014	0.222 ± 0.014	0.224 ± 0.015
Lyon	0.120 ± 0.013	0.127 ± 0.013	0.142 ± 0.013	0.152 ± 0.014	0.214 ± 0.015	0.308 ± 0.017	0.340 ± 0.015

BEZ	0.197 ± 0.014	0.263 ± 0.014	0.298 ± 0.015	0.328 ± 0.015	0.332 ± 0.015	0.336 ± 0.015	0.340 ± 0.015
MSS	0.194 ± 0.014	0.276 ± 0.014	0.314 ± 0.015	0.309 ± 0.016	0.315 ± 0.016	0.319 ± 0.015	0.319 ± 0.015
Spec-log	0.256 ± 0.016	0.312 ± 0.017	0.322 ± 0.016	0.336 ± 0.016	0.338 ± 0.016	0.319 ± 0.017	0.313 ± 0.016
Spec-log1plus	0.231 ± 0.015	0.253 ± 0.016	0.259 ± 0.016	0.257 ± 0.016	0.261 ± 0.015	0.258 ± 0.015	0.248 ± 0.015
Spec-power	0.239 ± 0.015	0.310 ± 0.017	0.328 ± 0.016	0.333 ± 0.016	0.336 ± 0.016	0.322 ± 0.016	0.317 ± 0.016
Spec-Hill	0.238 ± 0.015	0.310 ± 0.017	0.320 ± 0.016	0.332 ± 0.016	0.336 ± 0.016	0.315 ± 0.016	0.312 ± 0.016
Multi-fiber	0.207 ± 0.014	0.289 ± 0.014	0.318 ± 0.015	0.332 ± 0.015	0.333 ± 0.015	0.338 ± 0.015	
Multi-threshold	0.297 ± 0.015	0.335 ± 0.016	0.352 ± 0.016	0.364 ± 0.016	0.358 ± 0.016	0.344 ± 0.016	
Mean CC_{norm} of the models for the DRC dataset							
WSR	0.113 ± 0.018	0.138 ± 0.019	0.197 ± 0.019	0.323 ± 0.024	0.411 ± 0.025	0.400 ± 0.025	0.397 ± 0.026
Lyon	0.174 ± 0.014	0.185 ± 0.015	0.228 ± 0.017	0.264 ± 0.018	0.272 ± 0.020	0.302 ± 0.020	0.344 ± 0.020
BEZ	0.038 ± 0.012	0.057 ± 0.012	0.105 ± 0.013	0.165 ± 0.017	0.212 ± 0.019	0.250 ± 0.021	0.264 ± 0.023
MSS	0.037 ± 0.008	0.040 ± 0.010	0.072 ± 0.009	0.097 ± 0.012	0.182 ± 0.014	0.226 ± 0.017	0.251 ± 0.019
Spec-log	0.206 ± 0.017	0.250 ± 0.020	0.363 ± 0.019	0.412 ± 0.024	0.433 ± 0.025	0.400 ± 0.025	0.413 ± 0.026
Spec-log1plus	0.204 ± 0.017	0.248 ± 0.020	0.357 ± 0.019	0.409 ± 0.024	0.445 ± 0.025	0.402 ± 0.026	0.414 ± 0.025
Spec-power	0.213 ± 0.016	0.250 ± 0.019	0.342 ± 0.020	0.403 ± 0.023	0.433 ± 0.026	0.392 ± 0.024	0.418 ± 0.027
Spec-Hill	0.211 ± 0.016	0.249 ± 0.019	0.338 ± 0.020	0.393 ± 0.023	0.422 ± 0.025	0.375 ± 0.022	0.413 ± 0.027

Multi-fiber	0.041 ± 0.010	0.061 ± 0.010	0.086 ± 0.013	0.143 ± 0.013	0.251 ± 0.018	0.289 ± 0.021	
Multi-threshold	0.215 ± 0.015	0.255 ± 0.019	0.343 ± 0.019	0.384 ± 0.023	0.406 ± 0.025	0.373 ± 0.023	

Table S2.2: Mean ± SE CC_{norm} between cortical responses and predicted responses for the NRF model.

Number of frequency channels			
Models	**8**	**16**	**32**
Mean CC_{norm} of the models for the natural sound dataset 1			
WSR	0.469 ± 0.016	0.476 ± 0.016	0.489 ± 0.017
Lyon	0.484 ± 0.014	0.524 ± 0.014	0.635 ± 0.012
BEZ	0.634 ± 0.011	0.637 ± 0.009	0.651 ± 0.009
MSS	0.704 ± 0.010	0.718 ± 0.010	0.732 ± 0.010
Spec-log	0.723 ± 0.013	0.741 ± 0.013	0.725 ± 0.015
Spec-log1plus	0.676 ± 0.015	0.697 ± 0.014	0.699 ± 0.014
Spec-power	0.707 ± 0.013	0.740 ± 0.011	0.740 ± 0.012
Spec-Hill	0.719 ± 0.012	0.733 ± 0.012	0.739 ± 0.013
Multi-fiber	0.706 ± 0.011	0.744 ± 0.009	0.760 ± 0.010
Multi-threshold	0.761 ± 0.012	0.774 ± 0.011	**0.776 ± 0.011**
Mean CC_{norm} of the models for the natural sound dataset 2			
WSR	0.199 ± 0.015	0.205 ± 0.015	0.203 ± 0.016
Lyon	0.130 ± 0.012	0.141 ± 0.013	0.220 ± 0.015
BEZ	0.312 ± 0.016	0.339 ± 0.016	0.347 ± 0.016
MSS	0.314 ± 0.016	0.336 ± 0.016	0.328 ± 0.016
Spec-log	0.341 ± 0.018	0.355 ± 0.017	0.346 ± 0.017
Spec-log1plus	0.278 ± 0.016	0.263 ± 0.016	0.251 ± 0.016
Spec-power	0.353 ± 0.017	0.360 ± 0.017	0.351 ± 0.017
Spec-Hill	0.341 ± 0.017	0.333 ± 0.017	0.334 ± 0.017

Multi-fiber	0.333 ± 0.015	0.350 ± 0.015	0.351 ± 0.015
Multi-threshold	0.378 ± 0.017	0.357 ± 0.017	0.353 ± 0.017
Mean CC_{norm} of the models for the DRC dataset			
WSR	0.045 ± 0.014	0.085 ± 0.017	0.155 ± 0.022
Lyon	0.125 ± 0.016	0.123 ± 0.018	0.134 ± 0.019
BEZ	0.069 ± 0.012	0.146 ± 0.013	0.153 ± 0.016
MSS	0.058 ± 0.011	0.085 ± 0.014	0.160 ± 0.014
Spec-log	0.289 ± 0.017	0.348 ± 0.021	0.351 ± 0.022
Spec-log1plus	0.255 ± 0.024	0.297 ± 0.027	0.329 ± 0.026
Spec-power	0.232 ± 0.016	0.328 ± 0.020	0.352 ± 0.022
Spec-Hill	0.219 ± 0.014	0.313 ± 0.019	0.345 ± 0.023
Multi-fiber	0.094 ± 0.012	0.127 ± 0.017	0.237 ± 0.019
Multi-threshold	0.282 ± 0.022	0.343 ± 0.026	0.376 ± 0.027

Chapter 3 | A dynamic network model of temporal receptive fields in primary auditory cortex

This chapter has been previously published as and adapted from: Rahman M, Willmore BDM, King AJ, Harper NS (2019) A dynamic network of temporal receptive fields in the auditory cortex. PLOS Computational Biology 15: e1006618

3.1 Introduction

In the previous chapter we explored how the prediction performance of the various encoding models of auditory cortical neurons depend on the choice of the pre-processing models. In this chapter we will focus mainly on the cortical model. We will examine various strategies for building encoding models ranging from simple linear non-linear to neural network models. We will develop a new class of neural network models with units possessing various characteristics of biological neurons.

As discussed in the two previous chapters, the tuning properties of auditory neurons are commonly described by a spectro-temporal receptive field (STRF) model, which characterizes the linear dependence of the neural response on the sound spectrum at a range of latencies (Aertsen et al., 1980; Aertsen and Johannesma, 1981a; deCharms et al., 1998; Theunissen et al., 2000; Klein et al., 2000; Schnupp et al., 2001; Miller et al., 2002; Escabí and Schreiner, 2002; Linden et al., 2003; Fritz et al., 2003; Gill et al., 2006; Christianson et al., 2008; David et al., 2009; Gourévitch et al., 2009). A static non-linearity is often applied to the output from the linear STRF - this linear non-linear (LN) model estimates the responses of neurons significantly better than the linear estimate (Machens et al., 2004; Simoncelli et al., 2004).

Neurons in the auditory cortex show dependence on stimulus history over a few hundred milliseconds to a few seconds (Asari and Zador, 2009). While STRF models are somewhat successful in explaining the dependence of neural responses on the past few hundred milliseconds of stimulus history, these models do not show how this temporal aspect of receptive fields might be implemented biologically. Naively, the direct biological instantiation of the STRF model would involve auditory cortical neurons receiving inputs at a range of simple delays spanning out to few hundred milliseconds; however, this is biologically unrealistic. The onset latencies of neurons in the ventral division of the medial geniculate body (vMGB), which provides the primary ascending input to primary auditory cortex (A1), are typically less than ~14 ms in the ferret (Homma et al., 2017). In the cat, they are typically less than ~20 ms (Calford, 1983), in the mouse less than ~10 ms (Anderson and Linden, 2011), and in the marmoset less than ~40 ms (Bartlett and Wang, 2011). This is far less than the duration of stimulus history which influences the responses of A1 neurons. Although a few vMGB units do show onset latencies beyond this range (in cat up to 70 ms (Calford, 1983), in mouse up to 60 ms (Anderson and Linden, 2011), and in marmoset up to 300 ms (Bartlett and Wang, 2011)), this small fraction is unlikely to be the main determinant of the dependence of the neural response in A1 on stimulus history beyond a few tens of milliseconds. This is because there are other plausible mechanisms, such as the dynamic temporally-integrative properties intrinsic to neurons.

In this study, we set out to model the temporal aspect of auditory neurons' receptive field using a biologically motivated modelling framework. Using recordings from anesthetized ferrets, we asked how much of the dependence on recent stimulus history - the temporal receptive fields, of primary auditory cortical neurons can instead be explained by certain of these simple dynamical aspects of neurons – an approach more consistent with the known biology. To model a neuron's response,

we constructed a neural network model whose output is the weighted sum of the responses of multiple units, each of which resembles an LN model. However, the response of each unit of the network was modified in accordance with a dynamic firing-rate equation. This low-pass filters the unit's response (by convolving with an exponential decay impulse response), providing a simple exponentially decaying memory. This integrative characteristic can be related to the capacitance and resistance of nerve cell membranes, and the individual time constant of each unit can be related to the membrane time constant of real neurons (Dayan and Abbott, 2001). The dynamic aspect of our network can alternatively be interpreted in terms of other neurobiological dynamical phenomena, such as channel-based neural adaptation, short term synaptic plasticity, or recurrent network properties. We found that our biologically-motivated dynamical model can accurately capture the temporal receptive fields of neurons without the need for a wide span of latencies of input.

3.2 Results

The neural response dataset used in this study was recorded from the primary auditory cortical areas, A1 and AAF, of anesthetized ferrets in response to a diverse selection of natural sounds (20 stimuli of 5 s duration), including speech, animal vocalizations and environmental sounds (Harper et al., 2016; Willmore et al., 2016). In total, 73 single-unit neural recordings showed sensitivity to the sound stimuli (see Methods), as measured by their noise ratios, and these were analyzed in this study.

To probe the computations underlying stimulus integration, models were fitted to estimate the neural firing rate (averaged over 20 stimulus repeats) of each neuron as a function of the preceding sound stimulation (Fig 3.1). First, for all models, the sound stimuli were pre-processed to generate a cochleagram (described in detail in Chapter 2) that approximates the processing that occurs in the cochlea and auditory nerve. We have used the spec-log model from Chapter 2 because of its simplicity and ability to predict cortical responses well. Second, using responses to 16 of the natural stimuli, a parameterized model was trained to estimate the firing rate as a function of the preceding cochleagram. After parameters of the models were fitted to the data, their fit quality was tested on a held-out test set composed of the responses to the remaining 4 natural stimuli.

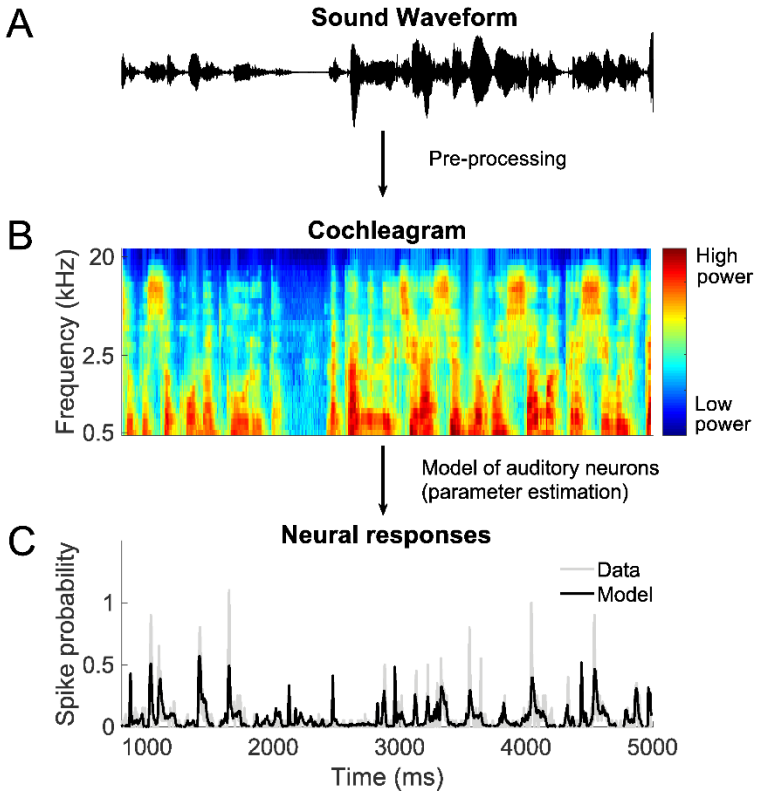

Figure 3.1: The models have two-stages. *A. The first stage involves pre-processing that mimics the auditory periphery, where the sound waveform is converted into a time-frequency representation (cochleagram). B. The output of the first stage is then used as input to a model that is fitted to the neural responses of auditory cortical neurons. C. After parameter estimation, the model was used to predict neural responses to novel stimuli.*

3.2.1 Stimulus history dependence of standard models

For each neuron, an LN model was fitted to estimate its firing rate as a function of the preceding cochleagram. The LN model consists of an STRF, followed by a sigmoid nonlinearity (Fig 3.2A). As mentioned in Chapter 1, the STRF is the weighted sum of the cochleagram over frequency components at a span of latencies. The effect of different spans of latencies (lengths) of the STRF was investigated. For the held-out test set, and averaged over the 73 neurons, the mean normalized correlation coefficient, mean CC_{norm} (Hsu et al., 2004; Schoppe et al., 2016), between the actual neural responses and the prediction of LN models with 25, 50, 100, 200, and 400 ms long STRFs are respectively 0.51, 0.64, 0.69, 0.71, and 0.71 (Fig 3.2B), where 1.0 is the maximum achievable prediction given the estimated neuronal noise and 0.0 indicates that there is no correlation between actual and predicted neural responses. The means of non-normalized Pearson correlation coefficients (CC) are 0.40, 0.50, 0.54, 0.55, and 0.55 respectively. Fig 3.2C shows STRFs of some example neurons estimated by the LN model.

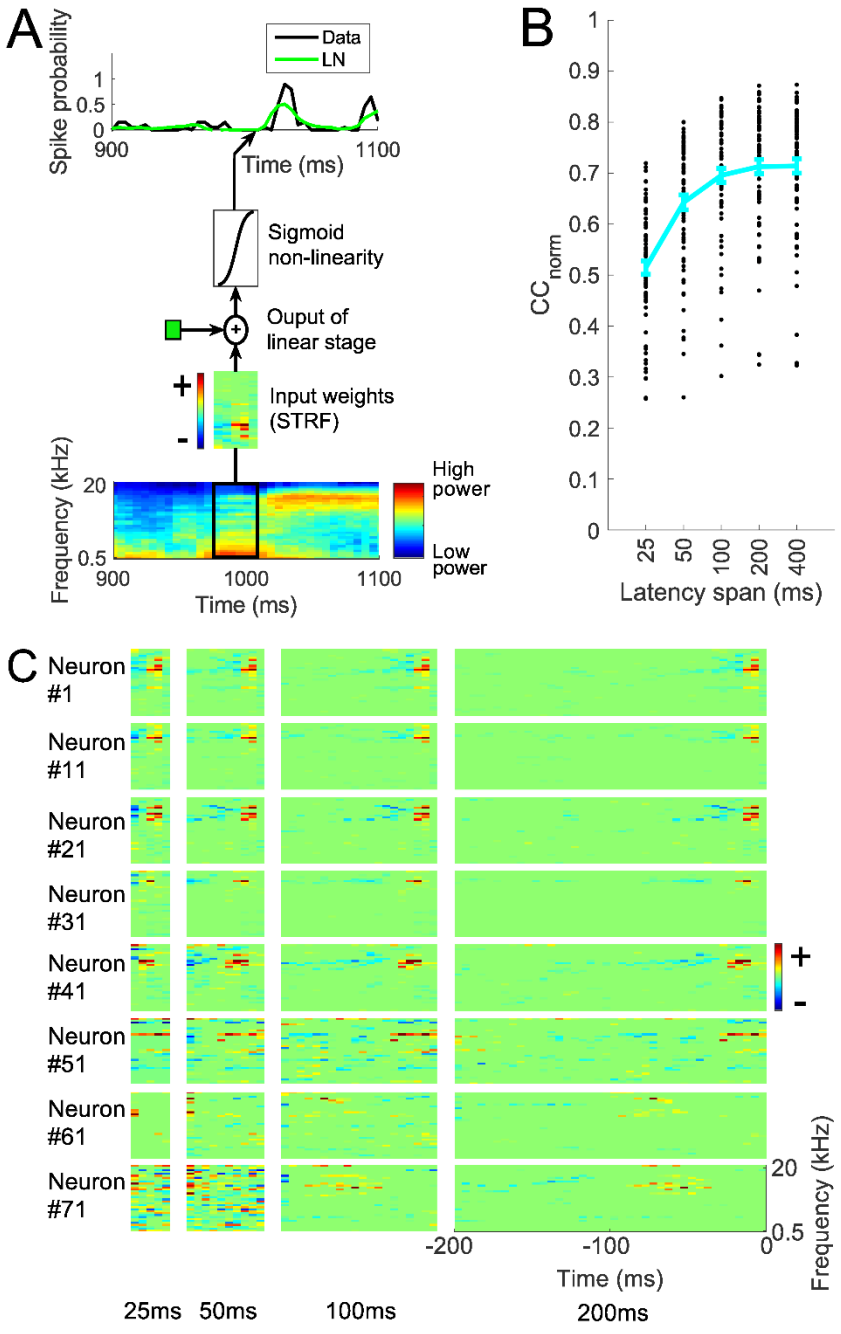

Figure 3.2: Prediction of neural responses by the linear non-linear (LN) model. *A. The LN model, which consists of a spectrotemporal receptive field (STRF) followed by a sigmoid nonlinearity. B. Normalized correlation coefficients (CC_{norm}) between neural responses and the prediction of the LN model with different latency spans. Each black dot represents a neuron, the curve indicates the mean over 73 neurons and error bars indicate standard error of the mean. C.*

STRFs of some example neurons, for LN models with different latency spans. Red, excitatory fields, blue, inhibitory. Excitatory and inhibitory fields extend up to 100 ms in most cases and up to 200 ms in some cases.

We also fitted each neuron to a network receptive field model (the NRF model) which is essentially an artificial neural network with a single hidden layer, using a sigmoid non-linearity in its hidden units and its single output unit (Fig 3.3A) (see Chapter 2 for more details). This model consisted of the weighted sum of multiple LN-like units. This model is very similar to the one published by Harper et al (2016), except that it uses sigmoid nonlinearities instead of a scaled tanh non-linearities. Having a sigmoid non-linearity makes this model consistent with the dynamic network model that we will come to, which requires non-negative outputs and hence also uses sigmoid non-linearities. Having multiple component STRFs in the NRF model allows for the modelling of neuronal sensitivity to non-linear interactions of multiple features (Harper et al., 2016). STRFs of at least 400 ms are needed for best performance in predicting neural responses (Fig 3.3B). For the test set, the mean CC_{norm} between neural responses and the model predictions made by NRF models with 25, 50, 100, 200 and 400 ms long component STRFs are 0.51, 0.64, 0.70, 0.72 and 0.73, respectively (mean CC are 0.40, 0.50, 0.54, 0.55 and 0.56, respectively). The NRF model predicts the neural responses equal to or better than the LN model over all durations examined, although by a more modest amount than seen in an earlier study (Harper et al., 2016). This may be because the NRF model in this study uses a sigmoid rather than tanh nonlinearity, or because the earlier study (Harper et al., 2016) used a different way of selecting the test set. Fig 3.3C shows the input weights (analogous to STRFs) for the 'effective' HUs (see Methods) of the model-fit for one example unit, obtained using the NRF model. Fig 3.3D and Fig 3.3E show the 'effectiveness' (see Methods) and IE score (see Methods) of the units shown in Fig 3.3C.

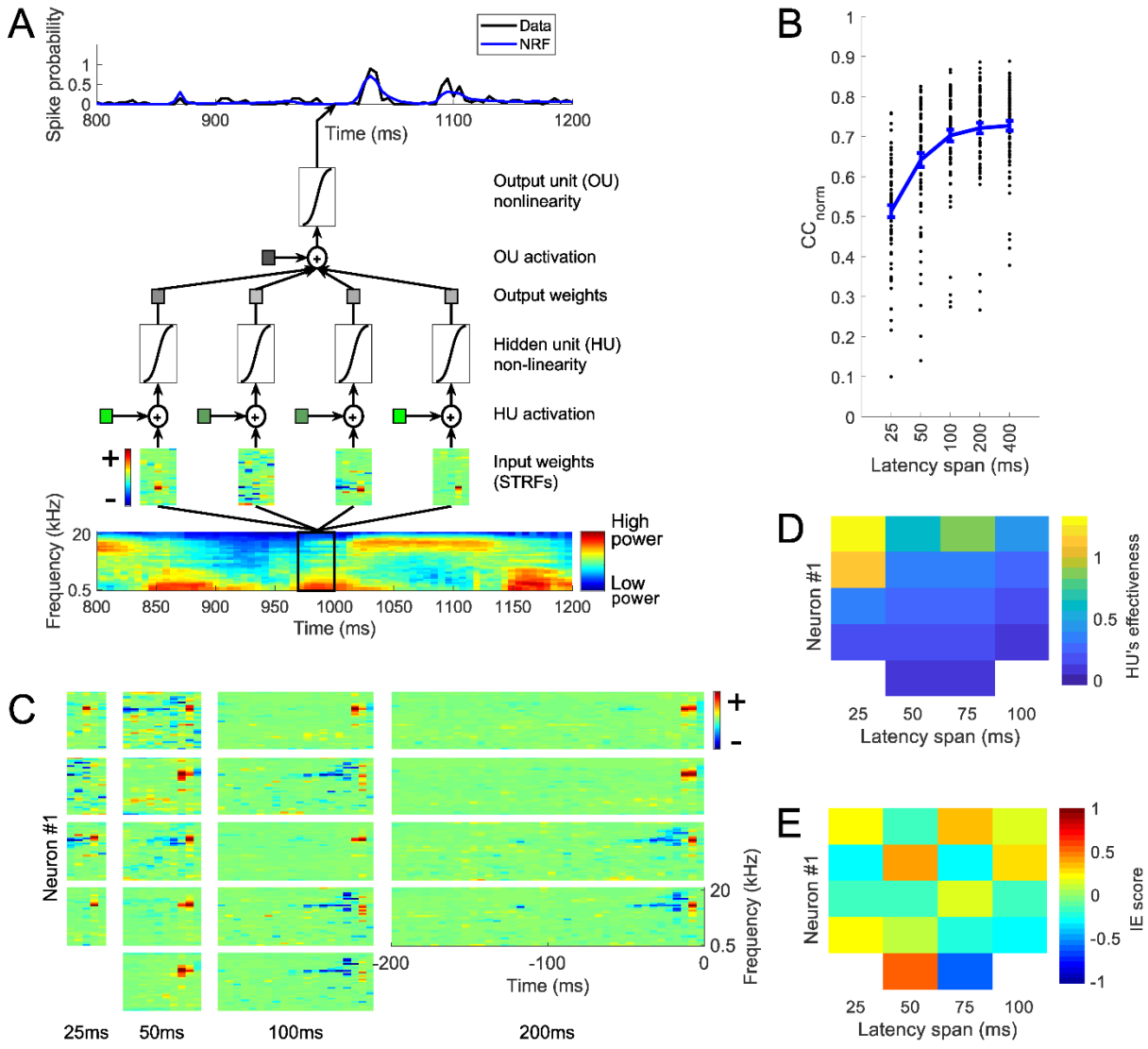

Figure 3.3: Prediction of neural responses by the network receptive field (NRF) model. *A. The NRF model, which involves a weighted sum of multiple LN-model like units, each with an STRF. B. Normalized correlation coefficients (CC_{norm}) between neural responses and the prediction of the NRF model with different latency spans. Each black dot represents a neuron, the curve indicates the mean over 73 neurons and error bars indicate standard error of the mean. C. Component STRFs of an example neuron. Each column represents an NRF model with different*

latency spans. D. 'Effectiveness' (see Methods) and E. IE score (see Methods) of the component STRFs shown in panel C.

3.2.2 Stimulus history dependence of a dynamic network model

Although both the LN and NRF models can capture dependence on stimulus history using a span of input latencies extending up to at least 200 ms, there is little biological evidence for such a wide span of latencies in the auditory thalamic neurons that feed to auditory cortex (see Introduction). With the aim of better understanding the biological underpinnings of stimulus history dependence in auditory cortical responses, a model with a dynamic aspect was developed that integrates the output of its component units over time in a manner similar to those of real neurons (see Methods). Each unit in the DNet model includes exponentially-decaying response integration, motivated by the integrative properties (Dayan and Abbott, 2001) of real neuronal membranes (Fig 3.4A). This model is in essence the NRF model but with dynamic units, each modelled by the dynamic firing-rate differential equation (see Methods, Equation 7) (Dayan and Abbott, 2001). That is, each unit of the model applies an exponentially decaying impulse response to the output of their non-linear activation function. Each unit's decay rate is governed by its individually fitted time constant, τ (see Methods).

We found that the mean CC_{norm} between the actual neural responses and the predictions of the DNet model with 25, 50, 100, 200 and 400 ms STRFs are respectively 0.71, 0.71, 0.70, 0.68 and 0.68 (Fig 3.4B) (mean CC are 0.55, 0.55, 0.54, 0.52 and 0.52, respectively). Because the 25-ms model performs the best and the biological range of latencies is typically less than 30 ms (Bizley et al., 2005), we focused on the 25-ms DNet model for further analysis. Figure 3.4C shows the input weights (analogous to STRFs) for the 'effective' hidden units (see Methods) of the model fits for 8 example neurons, using a DNet model with 25-ms long STRFs. Figure 3.4D and E show the 'effectiveness' (see Methods) and IE score (see Methods) of the units shown in Fig 3.4C.

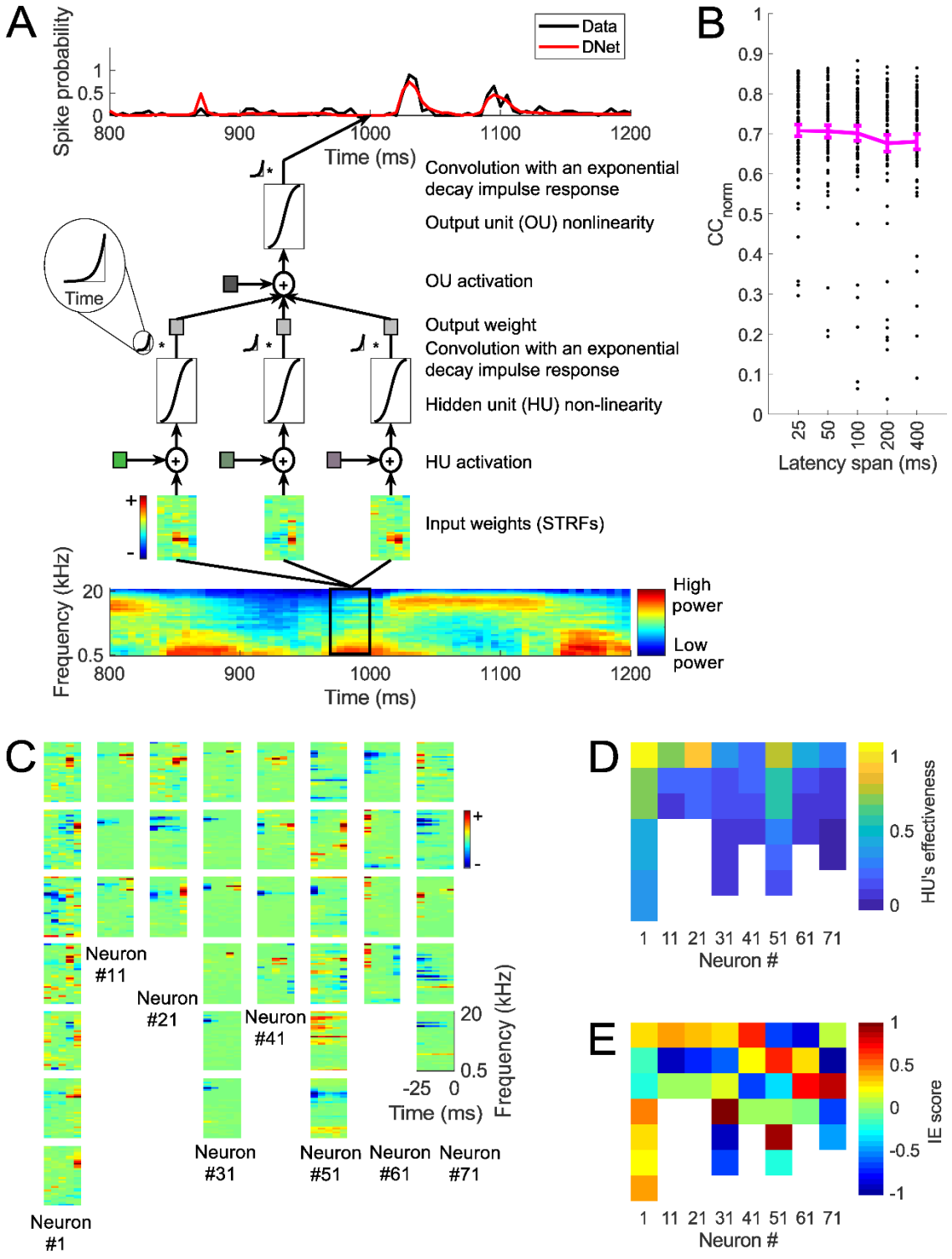

Figure 3.4: Prediction of neural responses using a dynamic network (DNet) model. *A. The DNet model. The cochleagram is passed through a set of linear-nonlinear filters, whose response is then integrated over time using an exponentially decaying impulse response to produce the*

hidden units' outputs. A weighted sum of these outputs is passed through a similar output unit to make a prediction of the neural response. B. CC_{norm} for the DNet model with different latency spans. Each black dot is a neuron, the curve indicates the mean over 73 neurons and error bars are standard error of the mean. C. The input weights (STRFs) for the 'effective' hidden units (see Methods) of 8 example neurons (each column is a different neuron) for the DNet model. D. 'Effectiveness' (see Methods) and E. IE score (see Methods) of the hidden units shown in panel C.

The 25-ms DNet model outperforms the DNet with longer (200 ms and 400 ms) STRFs, as measured by prediction of neural responses on a held-out test set averaged over all 73 neurons (Fig 3.5A). It also achieves performance similar to the best performing LN model, but performs a little worse than the best performing NRF model (Fig 3.5A). This can be seen qualitatively in the time course of the prediction for an example neuron and stimulus in Fig 3.5B. A neuron by neuron comparison shows that this result is consistent for most neurons (Fig 3.5C and 3.5D). Also, the 25-ms LN and NRF models are outperformed by a large margin by the 25-ms DNet model (Fig 3.5A). A neuron by neuron comparison shows that this result is also consistent for almost all neurons (Fig 3.5E and 3.5F).

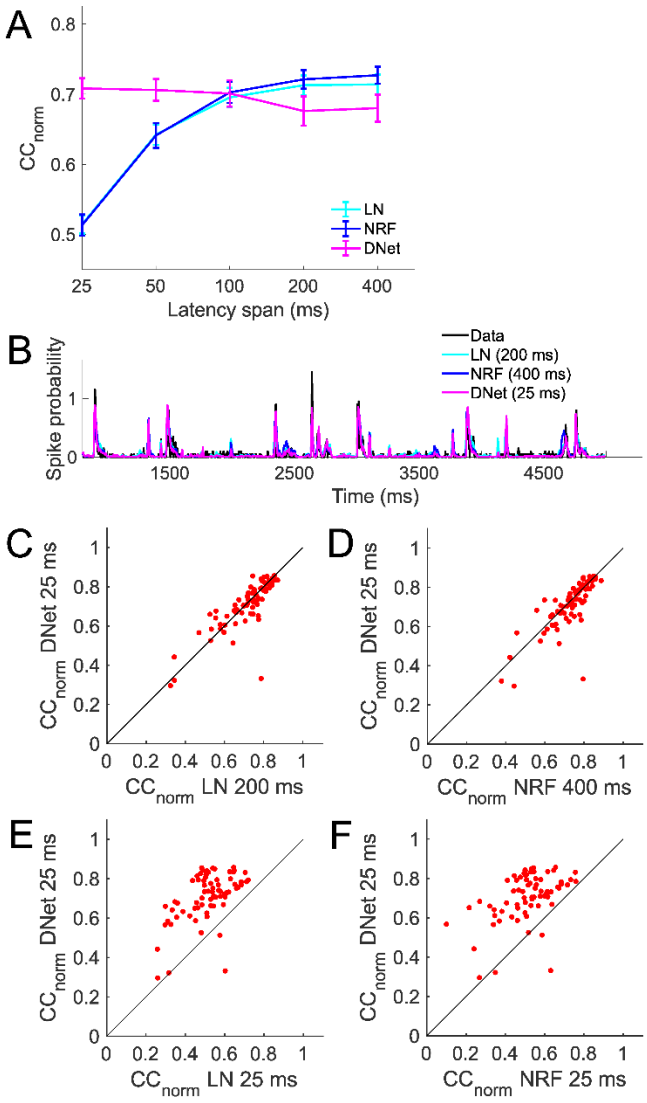

Figure 3.5: Comparison of the performance of the DNet model with other models. *A. Mean CC_{norm} of linear-nonlinear (LN), network receptive field (NRF) and dynamic network (DNet) models. B. Time course of predictions by best performing versions of each of the LN, NRF and DNet models, compared with real neural responses for an example neuron and stimulus. C. Neuron by neuron comparison of CC_{norm} values between the 25-ms DNet model and 200-ms LN model. Each red dot is a neuron. D. Neuron by neuron comparison of CC_{norm} values between the 25-ms DNet model and 400-ms NRF model. E. Neuron by neuron comparison of CC_{norm} values between the 25-ms DNet model and 25-ms LN model. F. Neuron by neuron comparison of CC_{norm} values between the 25-ms DNet model and 25-ms NRF model.*

Note that the DNet model is equivalent to the NRF model when the time constant for every DNet model unit is at the smallest possible value (d = 0). Hence, because all possible NRF model fits are a subset of all possible DNet model fits, the DNet should in theory always equal or outperform the NRF model. The fact that this is not the case suggests that at least in some cases the very best fits are not being found for the DNet model, perhaps as a consequence of some overfitting. That the DNet model performs worse for large latency spans, where there are more variables, is again probably due to challenges in fitting the DNet model, likely due to some overfitting to the training set.

3.2.3 Necessity of network structure and response integration for an effective DNet

To understand how much of an effect HUs of the DNet model with long time constants have on the model's predictions, the fitted model was modified by 'knocking out' (see Methods) HUs with time constants longer than the duration of their STRFs. This was achieved by setting the output weights of those units to zero, while keeping all other parameters intact. We found that the 25-ms DNet model is the DNet whose CC_{norm} is most reduced by this alteration (Fig 3.6A).

A single-unit dynamic model was also developed to examine the effect of a time constant in an LN-like single STRF model. Prediction performance of this model was found to be worse than the LN model (Fig 3.6B). This indicates that both long time constants and a network architecture are needed for the 25-ms DNet model to achieve prediction performance similar to that of standard models.

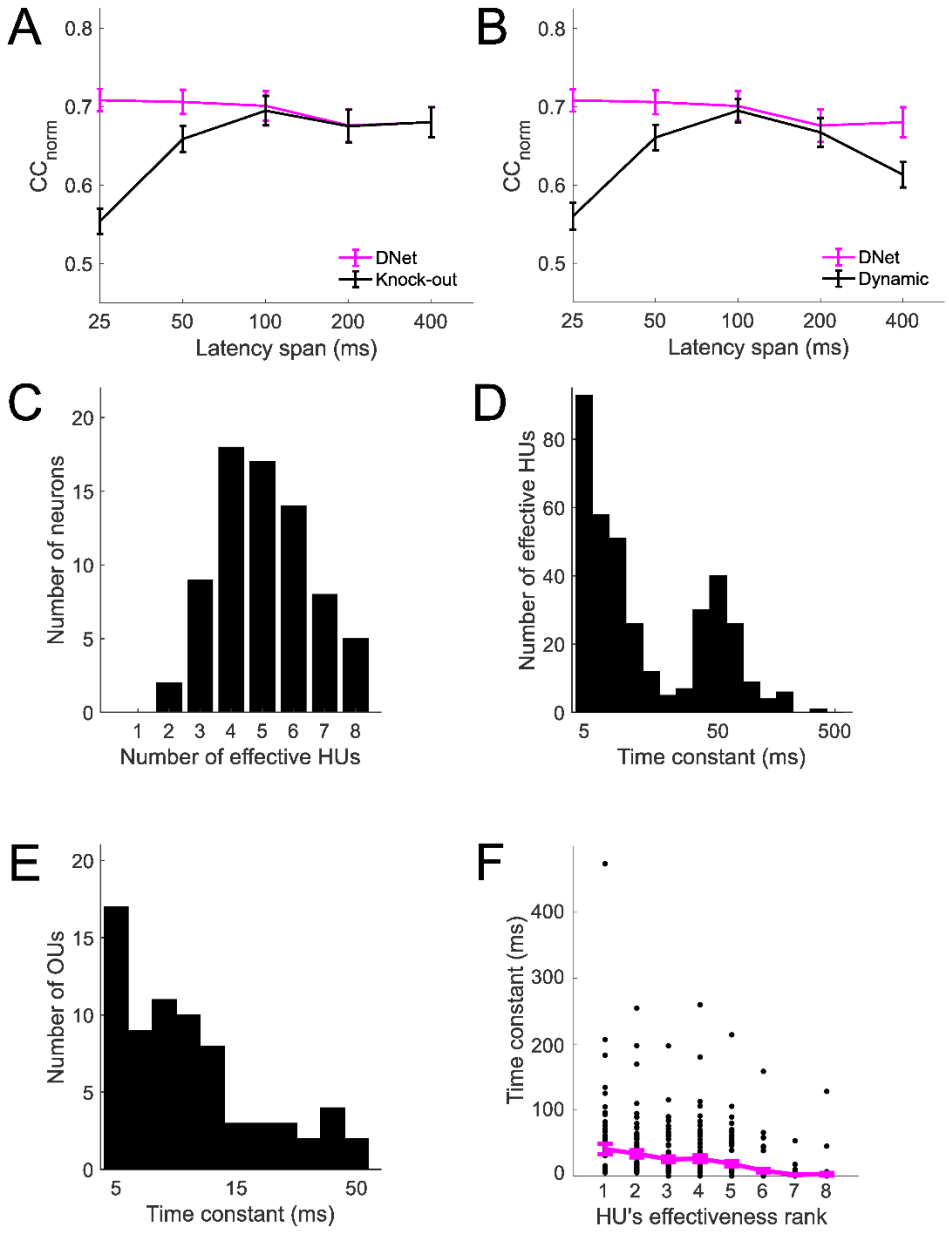

Figure 3.6: Multiple time constants in a network architecture are required to capture the dependence of neural response on stimulus history. *A. DNet model performance when large time constants are knocked out. B. Performance of the single-unit dynamic model. C. Number of effective hidden units (HUs) used to model each of the 73 neurons. D. Distribution of time constants of HUs. E. Distribution of time constants of the output units (OUs). F. Relationship between time constant and HU effectiveness.*

Although the DNet model had 20 hidden units, the L1 regularization eliminated unnecessary units. Hence there were fewer effective HUs, and for the 25-ms DNet model this number was found to be between 2 and 8 (Fig 3.6C). We defined an effective HU as one whose effectiveness (variance of weighted output, see Methods) exceeded a certain threshold.

In the 25-ms DNet model, the time constants of the effective HUs ranged between 5 and 475 ms, often far greater than the duration of the STRFs (Fig 3.6D). The time constants of the OUs of the DNet model ranged from 5 to 130 ms (Fig 3.6E). To investigate whether long time constants are associated with effective HUs, for each neuron, HUs were ranked in order of effectiveness and the mean of the time constants of all HUs in the same rank (over all neurons) was calculated. This showed that high-ranked HUs tend to have long time constants (Fig 3.6F).

3.2.4 Hidden units with long time constants in the DNet model

We identified the effective HUs with long time constants, i.e., where the value of time constant is greater than the length of the STRF (25 ms). To test whether a relationship exists between the time constant and unit's inhibitory/excitatory nature, we classified the effective HUs of the 25-ms DNet model into excitatory or inhibitory using an IE score (see Methods). If the output weight is positive, an IE score of 1 means a unit with entirely excitatory influence and an IE score of -1 means a unit with entirely inhibitory influence. We found that almost all units having large time constants are inhibitory in nature (Fig 3.7).

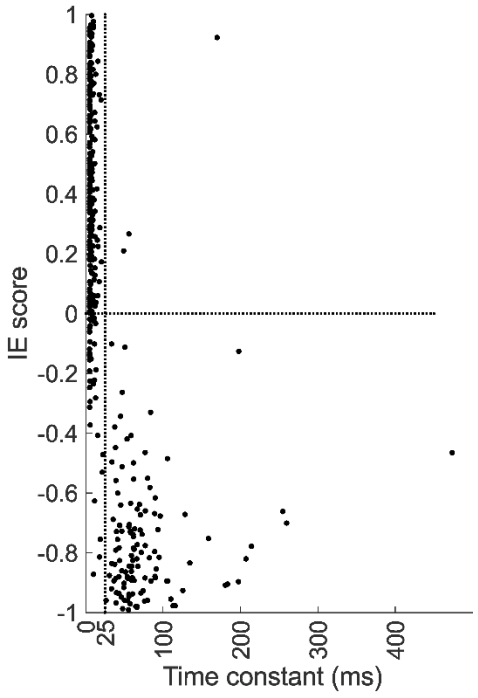

Figure 3.7: Relationship between inhibitory/excitatory (IE) score and time constant. *Almost all the HUs with time constants greater than the duration of the STRFs (units right of the vertical dotted line, the duration of the 25 ms STRFs) are inhibitory in nature. IE score = inhibitory/excitatory score.*

3.2.5 Membrane time constant versus synaptic time constant

In the DNet model, the integration governed by the time constant occurs after each unit's nonlinearity. Although the time constant is conceived and implemented in a manner consistent with a membrane constant, within the network it may capture other dynamic neural phenomena such as synaptic time constants, forms of a channel-based or synaptic adaptation, or certain forms of network recurrency. An alternative network model places the integration just before the nonlinearity, which is more consistent with the time constant being synaptic in nature (Dayan and Abbott, 2001). To investigate how much of the neural activity can be explained by such a model,

we modified the DNet model to form the sDNet model (synaptic DNet model), with each unit governed by Equations 13 and 14 (see Methods) (Dayan and Abbott, 2001).

The sDNet model achieves best performance in predicting neural responses when the duration of the STRFs is 200 ms (CC_{norm} = 0.71) (Fig 3.8A). Although the sDNet with 25-ms STRFs predicts neural responses better (CC_{norm} = 0.67) than the 25-ms LN model (CC_{norm} = 0.51) and the 25-ms NRF model (CC_{norm} = 0.51), it is worse than the 25-ms DNet model (CC_{norm} = 0.71). This suggests that a large percentage of the variance of neural responses can be explained by a time constant before non-linearity, but that having a time constant after the non-linearity further improves prediction.

While the largest time constants of the effective HUs of the sDNet model can be as large as ~470 ms (Fig 3.8B), the distribution of time constants of output units have smaller values compared to the DNet model (Fig 3.8C). As with the DNet model the most effective hidden units are those with longer time constants (Fig 3.8D) and units with long time constants are inhibitory in nature (Fig 3.8E).

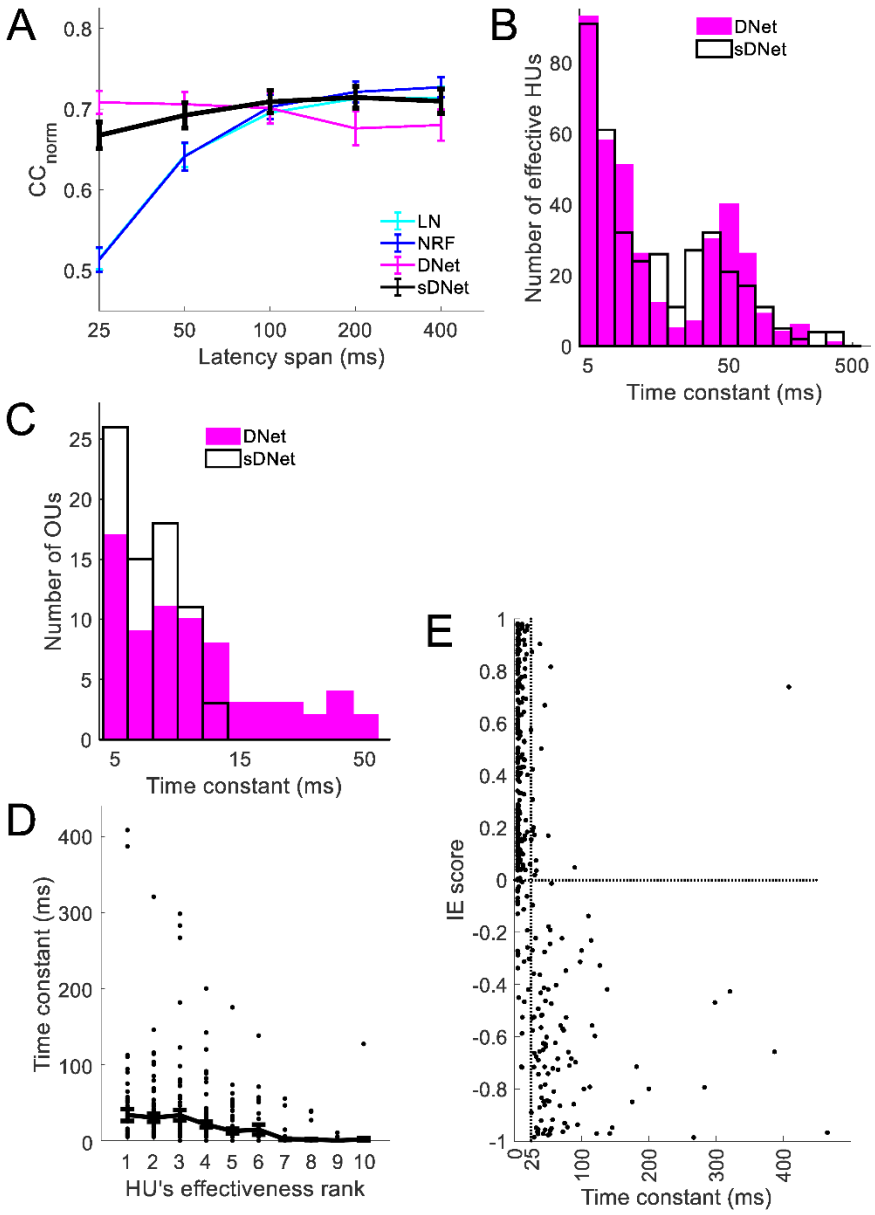

Figure 3.8: A modified DNet model with synaptic time constants (sDNet model). *A. Performance of the sDNet model. B. Distribution of the values of time constants of hidden units (HUs) in the sDNet and DNet model. C. Distribution of the values of time constants of the output units (OUs). D. Relationship between a HU's time constant and its effectiveness. E. Relationship between a HU's inhibitory/excitatory (IE) score and its time constant. Most of the HUs with time constants greater than the duration of the STRFs (units right of the vertical dotted line) are inhibitory in nature.*

3.3 Discussion

We find that a model of primary auditory cortical neurons that incorporates exponentially decaying memory processes – the DNet model – can capture some of the temporal aspects of the receptive field using very short (25 ms) STRFs in a network. Standard models such as the LN model (and the NRF model) represent this dependence using delay lines whose latencies extend up to a few hundred milliseconds. Our results suggest that, beyond 25ms, the temporal receptive field can be modelled more succinctly by exponentially-decaying memory processes than by delay lines.

When auditory neurons are fitted using a standard LN model, the resulting STRFs often contain narrowband inhibitory fields that decay over approximately 200 ms. In contrast, we find that the STRFs of the DNet model have short inhibitory fields with long time constants. This suggests that these temporally extended inhibitory fields in STRFs may reflect neural mechanisms that can be approximated by exponentially-decaying memory processes.

Although they have not been described in the ferret, a few vMGB units recording in other species do show onset latencies beyond a few tens of milliseconds: in cat latencies up to 70 ms have been described (Calford, 1983), in mouse up to 60 ms (Anderson and Linden, 2011), and in marmoset up to 300 ms (Bartlett and Wang, 2011). Areas other than vMGB that project to cortex may also include long-latency responses. Hence, we cannot rule out the possibility that long-latency inputs at least partially explain the dependence of auditory cortical neurons on stimulus history – indeed it is likely that they do to some extent. However, these long-latency responses are relatively rare, and we show that it is possible to explain a lot of the dependence of the neural response in A1 by exponentially-decaying memory properties consistent with those properties intrinsic to neurons, such as membrane dynamics.

There are many biological mechanisms which may underlie exponentially decaying memory processes. Neurons have various intrinsic dynamic integrative and adaptive characteristics originating from membrane dynamics, short-term synaptic plasticity, post-synaptic potentials, and channel-based adaptation (Dayan and Abbott, 2001). Recurrent network effects could also underlie such memory processes. In a previous study, a model with synaptic depression has been shown to partially capture the integration of stimulus history in A1 (David and Shamma, 2013). Here, we found that a better prediction of neural responses can be achieved by a model based on the integrative properties of neural membranes (the DNet model) than by the one based on the dynamics of synaptic potentials (the sDNet model). We note, however, that although the DNet model was developed with membrane dynamics in mind, aspects of the above listed dynamic processes could be captured by the same exponentially-decaying memory. These different processes are not exclusive of each other, and combinations of them may contribute to the effects we find.

In our study we have shown that adaptive properties that are exponential in nature could underlie neuron's temporal receptive fields. Although, radical alterations of the adaptation term could cause the model to perform very poorly, more subtle variations more consistent with the biology, for example, exploring the addition power law adaptation (Drew and Abbott, 2006), may perhaps improve the model performance. Identifying exactly which exact mechanism underlie the adaptation is challenging as there are so many non-exclusive possibilities each of which often differs only subtly. Careful investigation of the literature regarding the timescales or forms of the different mechanisms may be one method to distinguish mechanisms by incorporating these details into models and comparing them. Some possible directions for future experiments could

include intracellular and imaging data in model fitting to help identify cellular mechanisms underlying adaptive properties of neurons.

The firing rate of A1 neurons has been demonstrated to depend on stimulus history up to ~4 s in the past (Asari and Zador, 2009). This dependence cannot be fully captured by linear models such as STRFs. This indicates that there are some nonlinear dynamic processes that provide history dependence that cannot be captured either by linear or LN models or by the DNet model. Perhaps this is because the DNet model has limited capacity to capture temporal dynamics that are not exponential in nature.

Much of the relationship between stimulus and neural responses results from neurons being embedded in a network structure and from the nonlinear properties of the units in this network. From the auditory periphery to the A1, there are numerous feedforward, feedback and lateral connections. Many previous models of auditory cortical neurons have not explicitly represented this network character. However, some recent models have incorporated non-linear contextual effects (Ahrens et al., 2008; David et al., 2009; Calabrese et al., 2011; Rabinowitz et al., 2012; Willmore et al., 2016). Furthermore, a few models do explicitly incorporate multiple receptive fields (Sharpee et al., 2004, 2011; Atencio et al., 2008, 2009; McFarland et al., 2013; Sharpee, 2013; Kozlov and Gentner, 2016) or more extensive network structure (Harper et al., 2016), which perform better in predicting neural responses of A1/AAF neurons than STRF models (Atencio et al., 2008; Harper et al., 2016).

When modelling the dependence of neural responses on stimulus history, one approach is to include the history of a neuron's spikes and those of its recorded neighbors in a generalized linear

model (GLM) model (Pillow et al., 2008; Calabrese et al., 2011). However, unlike the DNet model, this approach does not model multi-layer network structure with hidden units that are inferred but not recorded. Modelling hidden network structure has been applied to the salamander retina, where in vitro responses to simple artificial stimuli are well predicted by a network model (LNFDSNF) with various non-linearities and response delays mimicking the retinal pathway (Real et al., 2017). The LNFDSNF model is particularly relevant to our study because it is a network model whose units have some memory of their past activity, though the form of this memory differs from that of the DNet model. The LNFDSNF model units have a feedback kernel that acts before the nonlinearity - in contrast, the DNet model units integrate their activity after the nonlinearity using an infinite impulse response that decays exponentially.

The succinct and biologically motivated description of neural responses provided by the DNet model provides a basis from which to develop further models to account for the ~30% of the CC_{norm} that remains to be explained. One variant of the model that can be developed involves a network with multiple hidden layers where the units each have just a single latency (i.e. each unit in the first hidden layer uses just a spectral receptive field at a single delay). This approach provides an opportunity to build deeper neural-network-like models with fewer parameters, potentially overcoming overfitting problems for deep network models of neural responses (Meyer et al., 2017). This network can provide a basis for new network models that incorporate a greater range of intrinsic dynamic processes more explicitly, for example modelling in each unit its membrane dynamics, synaptic depression and channel-based adaptation together, with different appropriate time constants in each case. Such models may show more predictive power.

Neurons are intrinsically dynamic systems, each with memory of past events. Furthermore, neurons are embedded in networks. We show that models incorporating both these characteristics can be valuable for explaining moment-to-moment *in vivo* responses of neurons in the auditory cortex. This is an arguably more biologically plausible way of capturing the history dependence of neural responses than a set of delay lines. In artificial neural networks, memory is usually implemented through recurrent connections between units. However, real biological neurons also have intrinsic memory processes with particular characteristics. Including such single-unit memory has potential advantages for explaining the *in vivo* responses of biological neurons and may also be useful in artificial neural networks. Hence, the value of our model is that it provides a biologically plausible alternative encoding model to the classic LN model, as well as a basis for further development of encoding models with hidden units with diverse intrinsic dynamic processes.

3.4 Methods

3.4.1 The dataset, noise ratio and cochleagrams

Stimuli

Models were fitted to single-unit neural responses of anesthetized ferrets to natural sound stimuli. Altogether, 20 sound clips were presented containing human speech in different languages, other animal vocalizations (e.g. ferret) and environmental sounds (e.g. wind and water). All clips were 5 s long with a sampling rate of 48,828.125 Hz and with root mean square intensity ranging from 75 to 82 dB SPL.

Experimental setup and neural responses

The electrophysiological data used in this study were taken from a series of experiments used in a previous study by Harper et al. (Harper et al., 2016). Briefly, recordings were made from the primary auditory cortical areas, A1 and the anterior auditory field (AAF) of 6 adult pigmented ferrets (5 females and 1 male) under ketamine (5 mg/kg/h) and medetomidine (0.022 mg/kg/h) anaesthesia for 20 repeats of the 20 natural sound clips played in random order. All animal procedures were performed under license from the United Kingdom Home Office and were approved by the local ethical review committee. In total, 56 penetrations resulted in 549 single and multi-units, of which 284 were single units. A single unit was taken for analysis only if its activity was driven by the stimulus according to the noise ratio (see below). For each unit, the number of spikes was counted in each 5 ms time bin and averaged over repeats, to provide a response profile $y_n(t)$, where t is time, and n is the clip number. The total number of time bins in a clip is T. For simpler notation, we will drop the subscript n, unless we note otherwise.

Noise ratio

The noise ratio was calculated as a measure of how much the response of each unit is dependent on the stimuli (Linden et al., 2003; Rabinowitz et al., 2011). The noise ratio was measured over all 20 stimuli and 20 repeats. Any units with a noise ratio > 40 were excluded from the study. Also, only putative single units were used. This resulted in 73 neurons for the study.

Cochleagrams

The models receive the input as a cochleagram, a spectrogram-like transformation of the sound waveform. The cochleagram approximates the spectral filtering performed by the auditory periphery and was calculated as follows (Harper et al., 2016; Willmore et al., 2016). For each sound clip, the amplitude spectrum was measured using 10-ms Hanning windows, overlapping by 5ms. The number of frequency channels was then reduced by weighted summation using overlapping triangular windows to provide 34 log-spaced channels (500 Hz to 22,627 Hz center frequencies, adapted from melbank.m (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html)). Next, a log function was applied to the value in each time-frequency bin, and any values below a low threshold were set to the threshold. The cochleagram was normalized to zero mean and unit variance over the training set (see Cross-validation and model testing). The input $x_{fq}(t)$ at time t was the chunk of cochleagram preceding t, given by $x_{fq}(t) = K_f(t - q + 1)$, where f is the frequency channel and q is the latency and $K_f(t)$ is the cochleagram.

3.4.2 The LN model

The two stage LN model consists of a linear model followed by a sigmoid output nonlinearity.

Linear stage:

$$a(t) = \sum_{f,q} w_{fq} x_{fq}(t) + b \ldots. (1)$$

where $a(t)$ is the model neuron's activation, w_{fq} is the input weight and b is the baseline activation. Both w and b are free parameters of the model and were estimated by regressing $y(t)$ against $x(t)$ using glmnet (Friedman et al., 2010). To overcome overfitting, the weights were regularized using an L1-norm (LASSO regularization). Thus $a(t)$ can be seen as the best linear estimate of $y(t)$ from $x(t)$. The regularization hyperparameter λ controlled the strength of the regularization. To select λ, the model was subjected to k-fold crossvalidation (see Cross-validation and model testing).

Nonlinear stage: The nonlinear stage of the model was a parameterized sigmoid nonlinearity:

$$v(t) = \frac{\rho_1}{1+\exp\left(\frac{-a(t)-\rho_3}{\rho_2}\right)} + \rho_4 \ldots (2)$$

The four parameters ρ_i of the function were fitted by minimizing the squared error between $v(t)$ and $y(t)$.

3.4.3 The Network Receptive Field (NRF) model

The NRF model is an artificial neural network with a single hidden layer of 20 hidden units (HU), which converge onto an output unit (OU). Each unit of the model is similar to a single LN model. The activation of the j-th HU, $a_j(t)$ is:

$$a_j(t) = \sum_{f,q} w_{jfq} x_{fq}(t) + b_j \ \ldots (3)$$

The output of the j-th HU, $v_j(t)$ is:

$$v_j(t) = \mathrm{g}\left(a_j(t)\right) \ldots (4)$$

where $\mathrm{g}\left(a_j(t)\right)$ is a sigmoid non-linear activation function, $1/(1 + \exp(-a_j(t)))$. The HUs feed these outputs to the OU. The activation function of the OU $a_o(t)$ is:

$$a_o(t) = \sum_j w_j v_j(t) + b_o \ \ldots (5)$$

The output $v_o(t)$ of the OU is:

$$v_o(t) = g\big(a_o(t)\big) \ \ldots (6)$$

which is the model's estimate of the neuron's response at time *t*.

3.4.4 The Dynamic Network (DNet) model

A neuron's response depends on the activity of its presynaptic neurons through the total synaptic current they evoke. However, a neuron's membrane potential, and consequently its firing rate, is not an instantaneous function of this current. Due to the capacitance and resistance of the neural cell membrane, the membrane potential and hence the firing rate is approximately a low-pass filtered version of the synaptic current. A common simple dynamic model of neurons that captures this phenomenon is the differential equation (Equation 7) known as the firing-rate equation (Dayan and Abbott, 2001).

$$\tau' \frac{dv'}{dt'} = -v' + \mathrm{F}\big(I(t')\big) \ldots (7)$$

Here, τ' is the membrane time constant, v' is post-synaptic firing rate, $I(t')$ is the synaptic current at time t', $F()$ is a nonlinear function that relates synaptic current to the post-synaptic firing rate, and t' is time. Here we use t', τ' and v' with a prime to denote that for this equation these variables operate in continuous time rather than discrete time. τ' determines how rapidly firing rate approaches steady state for a constant synaptic current, and can be related to the membrane time constant (Dayan and Abbott, 2001). Although the relationship is not strict, we shall call τ' the membrane time constant. This is equivalent to convolving the instantaneous output of function $F()$ with an exponential decay impulse response.

Equation 7 can be discretized by the Euler method, where $\Delta t'$ = 1 time-bin = 5 ms. Discretized time is expressed as t instead of t' and the discretized form of the firing rate equation is:

$$v(t) = \left(1 - \frac{1}{\tau}\right) v(t-1) + \frac{1}{\tau} \mathrm{F}\big(I(t)\big) \ldots (8)$$

The dynamic network model incorporates units that behave like Equation 8. In recognition of the fact that the characteristics of units in our model may not exactly correspond to biophysical properties of neurons described by the firing rate equation, and to be consistent with the NRF model, we rename $F()$ as $g()$ and the current $I(t)$ as activation $a(t)$.

As a result, activation $a_j(t)$ of the j-th HU of the DNet model is the same as the NRF model:

$$a_j(t) = \sum_{f,Q} w_{jfQ} x_{fQ}(t) + b_j \ldots (9)$$

However, the output $v_j(t)$ is different from that of the NRF model, incorporating the discretized firing-rate equation:

$$v_j(t) = \left(1 - \frac{1}{\tau_j}\right) v_j(t-1) + \frac{1}{\tau_j} \mathrm{g}\left(a_j(t)\right) \ldots (10)$$

Similarly, the activation $a_o(t)$ of the OU is:

$a_o(t) = \sum_j w_j v_j(t) + b_o$ … (11)

and the output $\mathrm{v}_o(t)$ of the OU is:

$\mathrm{v}_o(t) = \left(1 - \frac{1}{\tau_o}\right) v_o(t-1) + \frac{1}{\tau_o}\mathrm{g}\left(a_o(t)\right)$ … (12)

Here, τ_j and τ_o are the membrane time constants of j-th hidden unit and output unit, respectively.

3.4.5 The sDNet model

The firing-rate equation (Equation 7) assumes that the time constant that governs the relationship between the current and the firing rate is substantially larger than the time constant that governs the decay of post-synaptic potentials. However, if we assume the reverse, i.e., that the synaptic time constant is substantially larger, a different simple model of the relationship between the firing rate and the current becomes appropriate (Dayan and Abbott, 2001). This model is given by:

$\tau_s' \frac{dI}{dt'} = -I + z$ …….. (13)

$v = \mathrm{F}(I)$ …….. (14)

Here, I is synaptic current and v is post-synaptic firing rate, z is the immediate influence of the presynaptic spikes on the synaptic current, $\mathrm{F}()$ is the nonlinear function of synaptic current to post-synaptic firing rate, τ_s' is synaptic time constant, and t' is time.

Similar discretization, like that used for the DNet model (see above), can be used for Equation 13 to allow this to be used for the units of a neural network model. Again, replacing the synaptic current, I, with the activation a, the activation of hidden units of the sDNet model can be modelled as the following:

$$a_j(t) = \left(1 - \frac{1}{\tau_{sj}}\right) a_j(t-1) + \frac{1}{\tau_{sj}} \left(\sum_{f,Q} w_{jfQ} x_{fQ}(t) + b_j\right) \ldots (15)$$

And the hidden unit output, replacing $\mathrm{F}()$ with $\mathrm{g}()$, is given by:

$$v_j(t) = \mathrm{g}\left(a_j(t)\right) \ldots (16)$$

Hence, in the sDNet model, the activation of each HU, rather than the output of the nonlinearity, is convolved with an exponential decay impulse response.

Similarly, for the output units,

$$a_o(t) = \left(1 - \frac{1}{\tau_{so}}\right) a_o(t-1) + \frac{1}{\tau_{so}} \left(\sum_j w_j v_j(t) + b_o\right) \ldots (17)$$

and the output,

$$v_o(t) = \mathrm{g}\big(a_o(t)\big) \quad \ldots (18)$$

Here, τ_{sj} and τ_{so} are the synaptic time constants of j-th hidden unit and output unit respectively.

3.4.6 Training and testing network models

Objective function of the NRF and DNet models

The free parameters w_{jfq}, w_j, b_j, and b_o (also τ_j and $\tau_o$for the DNet and τ_{sj} and τ_{so} for the sDNet) were optimized by minimizing the squared error between $v_o(t)$ and $y(t)$ subject to L1-regularization of the weights. Thus, the objective function is given by:

$$E = \frac{1}{2NT}\sum_{n,t}\Big(v_{o,n}(t) - y_n(t)\Big)^2 + \lambda\big(\sum_{j,f,q}|w_{jfq}| + \sum_j|w_j|\big) \ldots (19)$$

Here, n is included to indicate clip number, but was left out of other equations for simplicity. N is the number of clips used in training.

Parameter estimation

All models except for the LN model are fitted by minimizing the objective function with respect to the free parameters using the sum-of-function optimizer algorithm (Sohl-Dickstein et al., 2013). In using this algorithm, we take one clip to be one minibatch. The optimization algorithm requires calculation of the error gradients in respect to each of the parameters. For the NRF model, error gradients are calculated using standard chain rule. For the dynamic models, the process is similar (see below).

For the network models, before training, the weights were initialized by modified Glorot initialization from a uniform distribution ranging from - $\frac{1}{\sqrt{fq+l}}$ to + $\frac{1}{\sqrt{fq+l}}$, where fq is the number of input weights to a HU and l = 1 is the number of output weights from a HU. The biases were initialized similarly (Glorot and Bengio, 2010).

For the DNet model, to prevent the $\frac{1}{\tau}$ term running into mathematical error (when $\tau = 0$), we substitute $\frac{1}{\tau}$ with $\mathrm{h}(d) = \frac{1}{1+d^2}$. Parameters, d_j and d_o (for hidden and output units) were initialized from the square root of an exponential distribution with a mean of 1. Similar, substitution is used for the sDNet model.

Error-gradient calculation for the DNet model

To estimate the parameters, the gradients of the error with respect to various parameters are calculated using a method similar to the one used for real time recurrent learning (Williams and Zipser, 1989).

Substituting $\frac{1}{\tau} = \frac{1}{1+d^2} = \mathrm{h}(d)$, the dynamic equations for HUs (Equation 10) and OU (Equation 12) of the DNet model respectively become:

$$v_j(t) = \big(1 - \mathrm{h}(d_j)\big) * v_j(t-1) + \mathrm{h}\big(d_j\big) * \mathrm{g}\Big(a_j(t)\Big) \ldots (20)$$

$$v_o(t) = \big(1 - \mathrm{h}(d_o)\big) * v_o(t-1) + \mathrm{h}(d_o) * \mathrm{g}\big(a_o(t)\big) \ldots (21)$$

Using the chain rule, the gradient of the error term, $\mathrm{E}(t)$ at time t in respect to the weights of the output layer, w_j,

$$\frac{d\mathrm{E}(t)}{dw_j} = \frac{d\mathrm{E}(t)}{dv_0(t)} * \frac{dv_0(t)}{dw_j} \ldots (22)$$

$$\frac{dv_0(t)}{dw_j} = \big(1 - \mathrm{h}(d_o)\big) * \frac{dv_0(t-1)}{dw_j} + \mathrm{h}(d_o) * \mathrm{g}'\big(a_o(t)\big) .* v_j(t) \ldots (23)$$

Similarly, the gradient of the error term, $\mathrm{E}(t)$ in respect to the bias of the output layer, b_o,

$$\frac{d\mathrm{E}(t)}{db_o} = \frac{d\mathrm{E}(t)}{dv_0(t)} * \frac{dv_0(t)}{db_o} \ldots (24)$$

$$\frac{dv_0(t)}{db_o} = \big(1 - \mathrm{h}(d_o)\big) * \frac{dv_0(t-1)}{db_o} + \mathrm{h}(d_o) * \mathrm{g}'\big(a_o(t)\big) \ldots (25)$$

Finally, the gradient of the error term, $\mathrm{E}(t)$ with respect to the parameter, d_o,

$$\frac{dE(t)}{dd_o} = \frac{dE(t)}{dv_0(t)} * \frac{dv_0(t)}{dd_o} \ldots (26)$$

$$\frac{dv_0(t)}{dd_o} = \big(1 - \mathrm{h}(d_o)\big) * \frac{dv_0(t-1)}{dd_o} - \mathrm{h}'(d_o) * v_0(t-1) + \mathrm{h}'(d_o) * \mathrm{g}\big(a_o(t)\big) \ldots (27)$$

Hence, the gradient is passed forward from the current time step to the next time step to be used in calculation of the gradient at that new time step. At $t = 1$, the values of $\frac{dv_0(t-1)}{dw_j}$, $\frac{dv_0(t-1)}{db_o}$ and $\frac{dv_0(t-1)}{dd_o}$ are undetermined. So, at time $t = 1$, the values of these terms are set to zero.

The gradients of the error term, $\mathrm{E}(t)$, with respect to the rest of the parameters of the DNet model and the sDNet model are obtained similarly.

Cross-validation and model testing

From the data for the 20 sound stimuli, 4 were chosen as a test set, which was not used during training and cross-validation. The cross-validation set (the remaining 16 stimuli) was used to fit the models using k-fold cross validation, where k = 8. The cross-validation set was randomly divided into a training set of 14 stimuli and a validation set of 2 stimuli. The model was trained on the training set for 18 different values of the hyperparameter λ. A log spaced range of lambda values was used, but with a somewhat lower density at the extremes. For the LN model, the exact values of λ used were: 1.00 x 10^{-1}, 2.00 x 10^{-2}, 1.17 x 10^{-2}, 6.84 x 10^{-3}, 4.00 x 10^{-3}, 2.34 x 10^{-3}, 1.37 x 10^{-3}, 8.00 x 10^{-4}, 4.68 x 10^{-4}, 2.74 x 10^{-4}, 1.60 x 10^{-4}, 9.36 x 10^{-5}, 5.41 x 10^{-5}, 3.20 x 10^{-5}, 6.40 x 10^{-6}, 1.28 x 10^{-6}, 2.56 x 10^{-7}, and 5.12 x 10^{-8}. For the rest of the models, the values of λ

used were: 1.00 x 10^{-3}, 2.00 x 10^{-4}, 1.17 x 10^{-4}, 6.84 x 10^{-5}, 4.00 x 10^{-5}, 2.34 x 10^{-5}, 1.37 x 10^{-5}, 8.00 x 10^{-6}, 4.68 x 10^{-6}, 2.74 x 10^{-6}, 1.60 x 10^{-6}, 9.36 x 10^{-7}, 5.41 x 10^{-7}, 3.20 x 10^{-7}, 6.40 x 10^{-8}, 1.28 x 10^{-8}, 2.56 x 10^{-9}, and 5.12 x 10^{-10}. For each of the fitted models, neural responses were then predicted for the validation set and the correlation coefficient between the actual neural responses and the prediction was measured. This process was repeated 8 times for different non-overlapping validation sets. The model was then retrained with the whole cross-validation set using the λ value that provided the highest mean correlation coefficient over all 8 folds. Next, the retrained network was used to predict the neural responses to the test set. All the correlation coefficients and normalized correlation coefficients shown are for this held-out test set, and all the model parameters shown are for the retrained network.

To ensure that all models receive the same amount of training data, models with different latency spans of STRFs are provided with exactly the same durations of neural response. This was done by clipping the necessary number of time points from the beginning of the neural response data.

3.4.7 Further analysis of the network models

Effective hidden units

The hidden layer of each of the networks contained 20 HUs, but because their weights were L1-regularized, they tended to develop substantive weights only if they explained aspects of the neural response that were not explained by any other units. As a result, many HUs developed weights close to zero, and thus the network models tended to have only a small number of effective HUs. The 'effectiveness' of each unit j was calculated (Harper et al., 2016) as the variance over time of the unit's weighted output $w_o v_j(t)$. Any HU with variance greater than 5% of the sum of the variances of all 20 HUs is considered effective.

IE score

The IE score (Harper et al., 2016) measures the degree to which a HU is excitatory or inhibitory, on a scale between -1 and 1. The most inhibitory is IE = -1, when all of the input weights are negative or zero, and the output weight is positive, or when all of the input weights are positive or zero and the output weight is negative. The most excitatory is IE = 1, when all of the input weights are positive or zero, and the output weight is positive, or when all of the input weights are negative or zero and the output weight is negative.

IE = sign(w_j) $\frac{\sum_{f,q} w_{jfq}}{\sum_{f,q} |w_{jfq}|}$ … (28)

Membrane time constant

Because the membrane time constant is modelled as $1 + d^2$ and the duration of each time step $\Delta t = 5$ ms, the value of the membrane time constant in ms is $5 * (1 + \mathrm{d}^2)$.

Knocking out large time constants of the DNet model

After fitting a network model, HUs with time constant greater than the length of the STRF are selected. The output weights connecting these units to the output unit are then set to zero. As a result, HUs with large time constants become inactive during the test process. Everything else in the model remained unchanged.

Chapter 4 | Decoding the past and estimating the future of sound inputs from the responses of auditory cortical neurons

4.1 Introduction

We have described in Chapter 3 how the response of auditory neurons is dependent on their stimulus history. Previous studies have shown that neural activity in A1 can be influenced by the stimulus past for up to several seconds (Asari and Zador, 2009). Some recent studies suggest that the auditory cortex is optimized to predict immediate future input (Singer et al., 2018). Here we explore to what extent the neural population response in the auditory cortex is predictive of future inputs as opposed to simply working as a 'sensory memory' of the past input. By decoding the past and future sound input from the neural responses, we examine the capacity of the auditory cortex to predict future sound inputs compared to encoding the recent stimulus history.

The brain uses sensory information to derive knowledge about the world and to make appropriate actions. Consequently, sensory signals that carry causal information about objects and events and guide action are the ones that are useful to the brain. So, a reasonable strategy for the brain could be to retain such information and discard the rest. One way to obtain information that might be useful for guiding action would be to filter in information that is predictive of the future state of the world (Bialek et al., 2001), and filter out non-predictive information. It has been proposed that the nervous system makes use of predictive information at all levels of its hierarchy (Palmer et al., 2013). Recent evidence suggests that this may happen even in the early stages of sensory processing (Palmer et al., 2013; Singer et al., 2018). By investigating the response of salamander retina to a moving bar Palmer et al. (2013) found that retinal ganglion cells operate optimally in

encoding the predictive information from the trajectory of the bar. Furthermore, they also measured to what extent spike trains in the retinal cells are predictive of their own spike history. The finding that spike trains optimally predict future spike suggests that the upstream brain areas might be encoding predictive information. Singer et al. (2018) took a modelling approach, in which, they predict future sound stimuli from the past history of same sound clip using a neural network model. In doing so, they discover that the receptive fields of the model match those found in auditory cortex. However, none of the prior studies have directly estimated from a population of neurons in the auditory cortex the ability of those neurons to predict future stimuli.

Decoding provides us with a way to study neural representations without making prior assumptions about the underlying encoding function. By projecting the neural responses back to the stimulus domain, decoding allows for a stimulus to be reconstructed in its full context as no explicit assumptions are made about which features of a sound stimulus are encoded. Previous studies have employed decoding to reconstruct sound spectrograms and features of speech from neural responses (Mesgarani et al., 2009; Mesgarani and Chang, 2012; Rabinowitz et al., 2013).

In the current study, we investigated the responses to diverse natural sounds of the neural population in ferret A1 (described in previous chapters). From this neural population response, we decoded the stimulus cochleagram using a linear decoder. Using a window of neural response immediately before the present, we ‘decoded’ both the stimulus that occurred before the present (the past stimulus) and the stimulus that occurred after the present (the future stimulus). That is, we compared how much of the past stimulus cochleagram can be estimated from auditory neural responses with how much of the future stimulus cochleagram can be predicted. Hence, we used the accuracy of this decoding as a quantification of the ability of the neural population to represent the past stimulus and predictions of the future stimulus. We also compared these measurements against the estimation of the past and future by a simple (linear-nonlinear-Poisson) encoding

model of auditory cortex. This was done to determine whether real auditory cortex contained additional non-linearities that enables it to better estimate past or future sensory stimulation.

4.2 Results

4.2.1 A linear decoder for estimating the past and future input from the neural population response

A linear decoder is often used to transform the population response into the original stimulus or some transformed version of the stimulus. In this case, we decode the cochleagram of sound from the recorded neural responses. The form of cochleagram that we used here is obtained by short-time Fourier transform with a small time bin (in this case 5 ms) of the sound signal, followed by taking the magnitude for each time bin and frequency, and then averaging over frequency bands using a set of triangular filters to provide 34 log-spaced channels (500 Hz to 22,627 Hz center frequencies). This time-frequency representation is then converted into common logarithmic values, which is passed through a threshold non-linearity to result in the cochleagrams $S(f,t)$ for subsequent analysis. Frequency channel index f goes from 1 to F = 34, and the time bin index t goes from 1 to T = number of time bins in the cochleagram. This is the same form as the spec-log model of Chapter 2, and exactly the same as the cochleagram model of Chapter 3.

The functional form of a decoder (Mesgarani et al., 2009) that reconstructs a frequency channel of a stimulus spectrogram or cochleagram from the neural population response is be given by Equation 1 below. We will use the notation of (Mesgarani et al., 2009) here for consistency with previous literature.

$$\hat{S}_f(t) = \sum_{n=1}^{N} \sum_{\tau=0}^{\Gamma-1} g_f(\tau, n) R(t - \tau, n) \ldots (1)$$

Here, $\hat{S}_f(t)$ the decoder's estimate of $S_f(t)$, where $S_f(t)$ is activity of frequency channel f of cochleagram $S(f,t)$. Also, $R(t,n)$ is the neural population response at time t and neuron n, where

n goes from 1 to N neurons. Finally, $g_f(\tau, n)$ is the decoding function, which is the weights for a linear mapping from the neural population response $R(t, n)$, at a span of lags τ= 0 to $\Gamma - 1$, to the estimate $\hat{S}_f(t)$ of the cochleagram. Γ is the span, in times bins, of the window of neural response that is decoded from.

We modified this linear decoder to provide an estimate $\hat{S}_{f,d}(t)$, of the past and future stimulus cochleagram at particular temporal 'deviation' d from the present, that is to estimate $S_f(t - d)$. By changing the value of the deviation, we can estimate a particular time point in either the past or the future.

$$\hat{S}_{f,d}(t) = \sum_{n=1}^{N} \sum_{\tau=0}^{\Gamma-1} g_{f,d}(\tau, n) R(t - \tau, n) \ldots (2)$$

Here, $g_{f,d}(\tau, n)$ is the decoding function that provides a linear mapping from the neural population response $R(t, n)$, at a span of lags τ= 0 to $\Gamma - 1$, to the estimate $\hat{S}_f(t)$ of the cochleagram for frequency channel f and deviation d, t is time in 5ms bins, and n indexes the neurons in the population. A negative time deviation predicts the future stimulus, a positive time deviation estimates the past stimulus.

For this study we used neural response dataset containing a population of 86 neurons. These are neurons that were recorded from anesthetized ferrets during the presentation of 20 natural sound clips (see Methods, also see natural sound dataset 1, Chapter 2, and Chapter 3). For each clip we produced model cochleagrams and took the trial-averaged spike-count responses of each neuron in 5 ms time bins.

We fit the linear decoder (Equation 2) to estimate the cochleagrams (of 16 of the clips), from the corresponding neural population responses (of the 86 neurons), using a given sliding temporal window of the neural response that is 5,10,…, or 200 ms wide (Γ = 1, 2, .. 40 in the units of time bins). We estimated the cochleagram up to 1000 ms in the past and up to 1000 ms into the future, relative to the time of the neural response window (the present). Figure 4.1 shows an illustration of this approach. After the linear decoder was trained on the data, the fitted parameters were then used to estimate the cochleagrams of the 4 novel clips not used in training, using the corresponding neural population responses as input. A *k*-fold training and testing procedure was used so all 20 clips could be used as novel clips for measuring estimation performance (see Methods) - all results shown are for novel clips and responses.

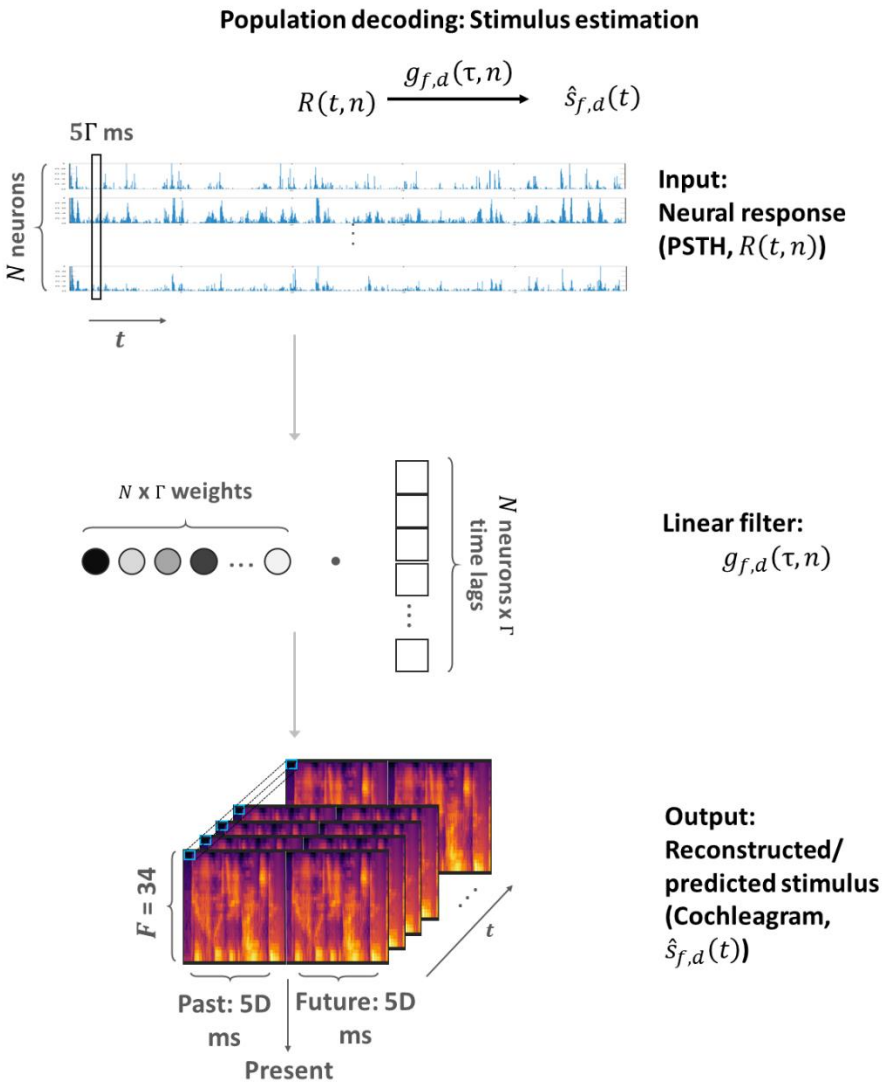

Figure 4.1: Illustration of stimulus decoding by the linear decoder. *A linear filter, $g_{f,d}(\tau,n)$, maps the neural population response, $R(t,n)$, into an estimate $\hat{S}_{f,d}(t)$ of the stimulus cochleagram $S_{f,d}(t)$. The stimulus cochleagram component at d time bins deviated from the present, $\hat{S}_{f,d}(t)$, is reconstructed from Γ time bins of population response. Using this method, we can reconstruct each frequency component f at different time points d into the past and future (up to D time bins).*

Figure 4.2 shows some of the cochleagram estimates for the test set (see Methods) of the decoder. The linear decoder estimates for past and future are plotted side by side, together with the actual cochleagram above the estimated ones for comparison. We also show examples of decoding from different window lengths of the neural response. It is clear from Figure 4.2 that the estimated cochleagrams more closely resemble the actual ones when the duration of the decoding window is longer (note how the estimation improves from the top to the bottom panels). We also observe that the estimated cochleagrams become dissimilar to the actual ones when the estimation is done at a far-off time-deviation from the present (the estimation becomes worse in moving from the left panel to the right panels). Estimation of future inputs is generally worse than decoding past inputs (Fig. 4.2).

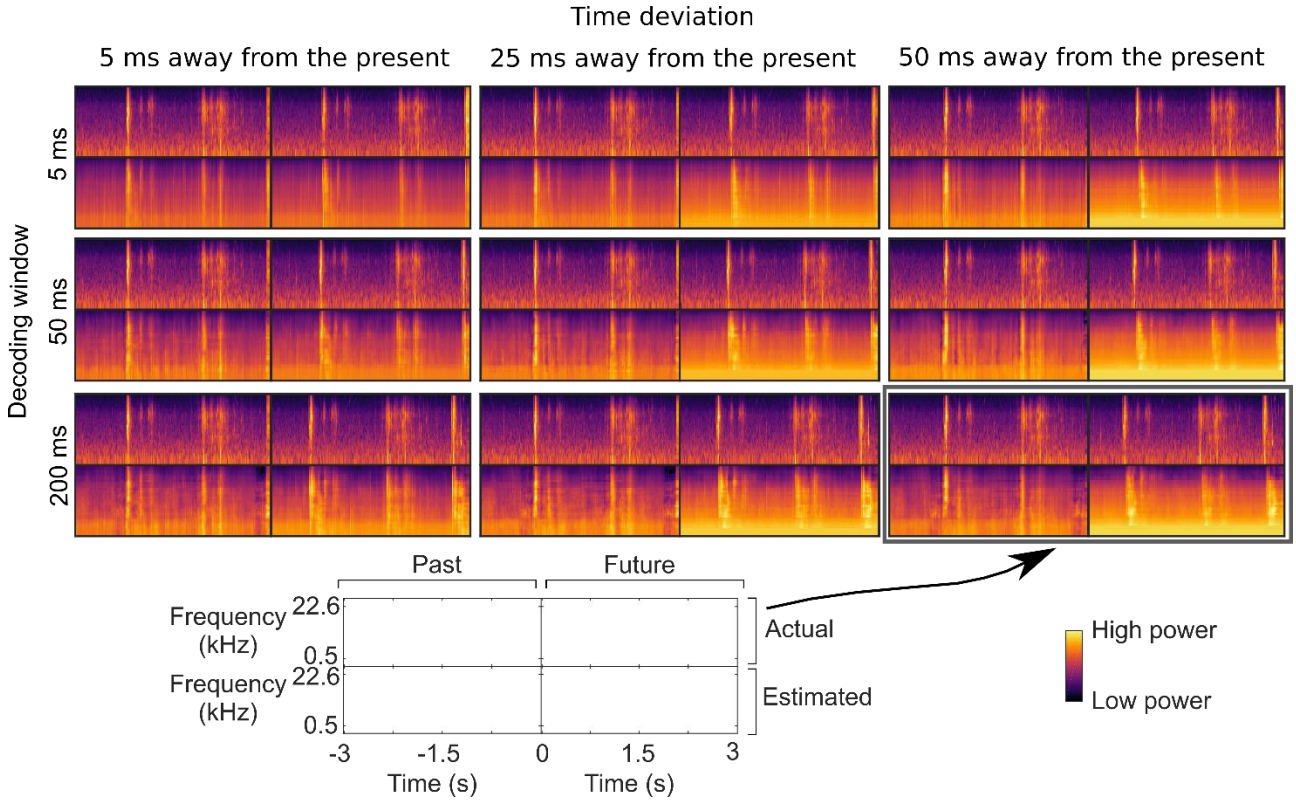

Figure 4.2: Examples of estimated stimulus cochleagrams. *Each panel shows estimates of a snippet of cochleagram for a novel sound clip of running water, when estimating into both the past and the future from A1 responses. The schematic at the bottom explains the organization of each panel. Panels in each row show the result of decoding from different neural response window widths and those in each column show decoding at various time deviations for both past and future inputs. The color scale in each panel ranges from the minimum to maximum value of each panel.*

4.2.2 Estimation accuracy for past and future input

To examine how well the stimulus can be decoded from the neural response, we assessed the accuracy of stimulus estimation using the correlation coefficient (CC) between the estimated

cochleagram and the actual cochleagram (for novel clips). The CC value for a decoder fitted for a particular time-point in the past/future and frequency channel, indicates the degree of similarity between the estimated and actual power in that frequency channel at that specific lag/lead relative to the present. The closer the value of CC to 1, the more accurate the estimate is. However, neurons are inherently noisy, so it is unrealistic to achieve a CC value of 1, or 100% accuracy.

As we mentioned, we followed a k-fold procedure to assess the performance of the decoder in estimating stimulus cochleagrams (see Methods for details). That is, we split the whole dataset of 20 stimulus clips into a training set (16 clips) and a test set (4 clips) five times to create five non-overlapping training and test set pairs. We trained the models for each training set, and estimated the cochleagrams from neural responses for the corresponding test set. We then measured the average CC values between the actual cochleagrams and estimated cochleagrams across all five test sets for different time and frequency points in the past and future (Fig. 4.3A). In a few instances, the CC at a given time-frequency point was undefined. This happens when the estimation of the decoder is always zero. We set the undefined CC values for these instances to zero. We used average CC values for different time and frequency points as an indicator of the extent to which the neural population response can be used to estimate the stimulus past and future.

We gain insights into various coding properties of the neural population from this analysis. Based on the estimation accuracy at various timepoints into the past and future in over all spectral channels, we found that it was possible to reconstruct the natural stimuli to some degree (CC >0.1) for up to a few hundred milliseconds of past stimulus and one or two hundred milliseconds of future stimulus (Fig. 4.3A). Although there is variation among frequency channels in the estimation accuracy, in general, estimation accuracy is consistent across frequency channels for up to ~700 ms into the past and up to 200 ms into the future. Increasing the decoding window of

the neural response makes the estimation of the past better, but this effect is less evident for decoding the future. When we take the average over frequency, estimation accuracy is very high close to the present, which decays and reaches a baseline after ~700 ms of the past and ~200 ms into the future (Fig. 4.3B). Beyond this time span, estimates for all spectral channels fall to a very low value (CC ~0.1). This remains little changed even when we shift the decoded past to account for different durations of decoding window (Fig. 4.3C).

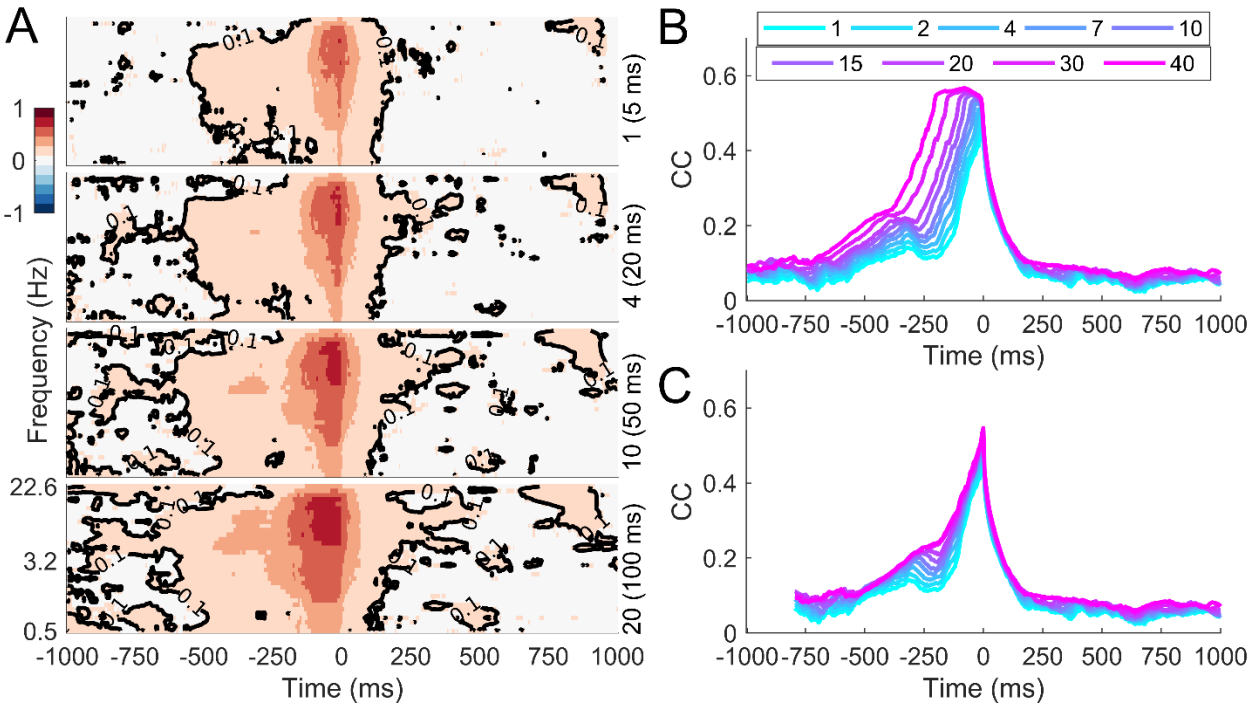

Figure 4.3: Accuracy of stimulus estimation from neural responses. *A. Estimation accuracy (correlation coefficient, CC) is shown across different frequency channels and time delays from the present to the past and the future. Each colored region in this figure represents the CC value. The black line shows the contour for a CC value of 0.1. Outside of this contour, CC values in the grey area fall between 0 and 0.1. Each plot in a different row represents CCs for decoding from different window length of neural response. Note that, decoding window lengths are given both in the units of time bins and ms. B. CCs averaged over frequency channels at various delays and for different lengths of the decoding window. Different colored lines represent average CCs for*

different decoding window lengths. Note that the length of the decoding windows is in the units of time bins (of 5ms). C. CCs averaged over frequency channels at various delays - in this case the past has been shifted away from the present to remove all the past stimulus bins that coincide with the decoding window of the neural response.

To investigate the effect of the length of decoding window on the accuracy of stimulus estimation, we examined how the average of all the CC values in Figure 4.3C up to 200 ms into the past or future changes with increasing the length of the decoding window (Fig. 4.4A). We found that increasing the length of decoding window causes more of an improvement in the estimation accuracy of the past stimuli, which continues increasing out to 200ms window length, than for that of the future stimuli, which saturates at around 75ms window length. This also holds when we examined how frequency-averaged CC changes with decoding window length relative to the 5 ms decoding window, at various time points into the past and the future (Fig. 4.4B). However, this effect of window length in improving the accuracy of stimulus estimation appears to saturate when the decoding window is ~75-150 ms long, with the exception of 100 and 200 ms into the past, which continues to improve with increasing length of decoding window out to 200 ms length (Fig. 4.4B).

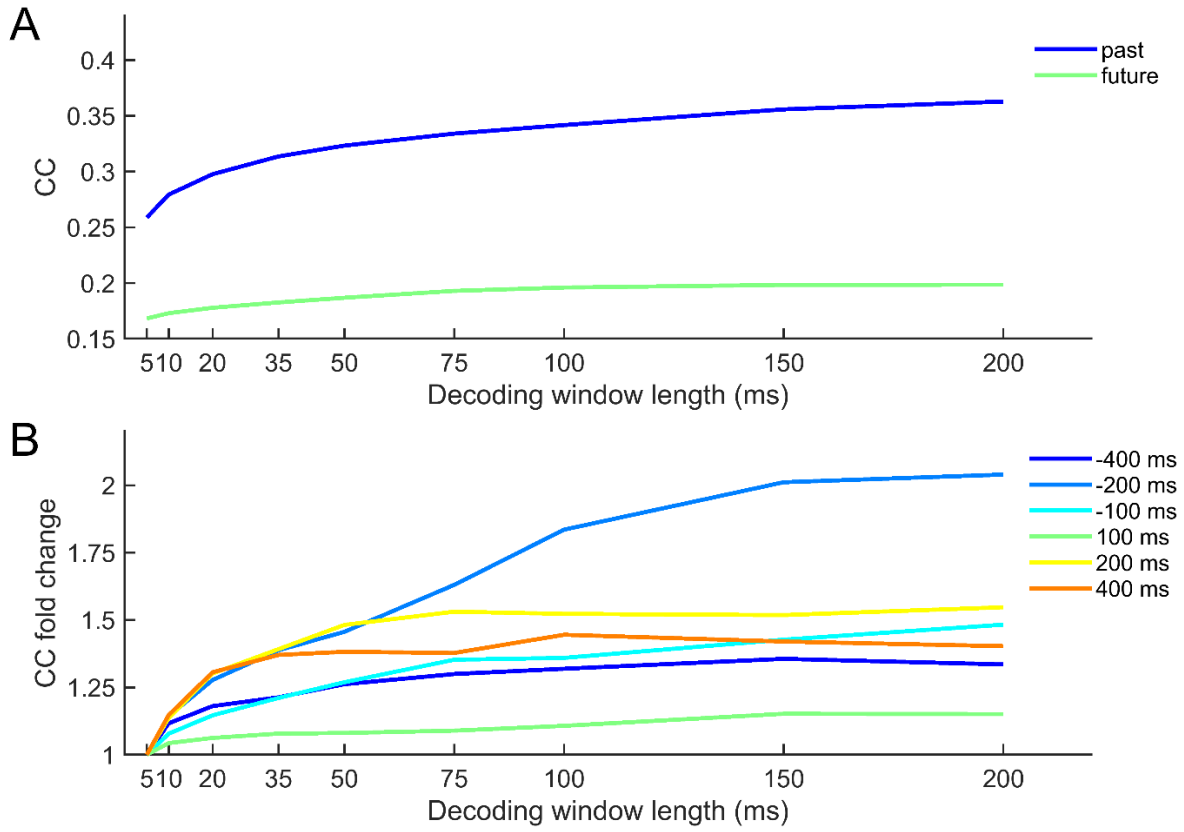

Figure 4.4: The effect of increasing the length of the decoding window on the accuracy of stimulus estimation. *A. Average of CC values up to 200 ms into the past and future for different decoding window lengths. Unlike Fig. 4.3B and C, we have shown the length of decoding window in the units of ms here. B. CC for different lengths of decoding windows, relative to the 5 ms decoding window, at various points into the past (at 100 ms, 200 ms and 400 ms away from the present) and the future (at 100 ms, 200 ms and 400 ms away from the present). Different colored lines represent CCs at various points into the past (-400 ms, -200 ms, -100 ms) and future (100 ms, 200 ms, 400 ms). For both panel A and B, the past has been time shifted to remove the stimulus bins that coincide with the neural response decoding window.*

4.2.3 Explaining the capacity of the neural population to encode the past and predict the future.

So far, we found that a neural population can to some degree encode stimulus past and predict stimulus future over the scale of a few hundred milliseconds. However, to understand the potential significance of these findings, we need to examine this in the context of a standard encoding model. We can ask whether the capacity to estimate the past and future of stimulus input from the neural response is the result of computations such as those present in a linear non-linear encoding model. To investigate this, we generated artificial neural responses using a linear-nonlinear-Poission (LNP) spiking model and then performed the decoding from those artificial responses. We fitted an LN model for each neuron in the population and then used the artificial response of the population to estimate the stimulus cochleagram. This shows us some interesting results. For capturing the past stimulus information, the LNP encoding model does better than neurons at a population level (Fig. 4.5). However, for predicting the future in low frequencies, the neural population does better than the responses of the LNP model. This is consistent across different decoding window durations. However, the capacity of real A1 neurons to predict the future better than the LNP model does not improve when the decoding window length is increased above 10 ms. Two possible interpretations come out of all this. Firstly, neurons might possess additional non-linearities on top of LN nonlinearity that help them to predict the future, perhaps at the cost of encoding the past less well. Secondly, the effect of such non-linearities on the prediction of future inputs by A1 neurons might be represented in a small time-window of their response.

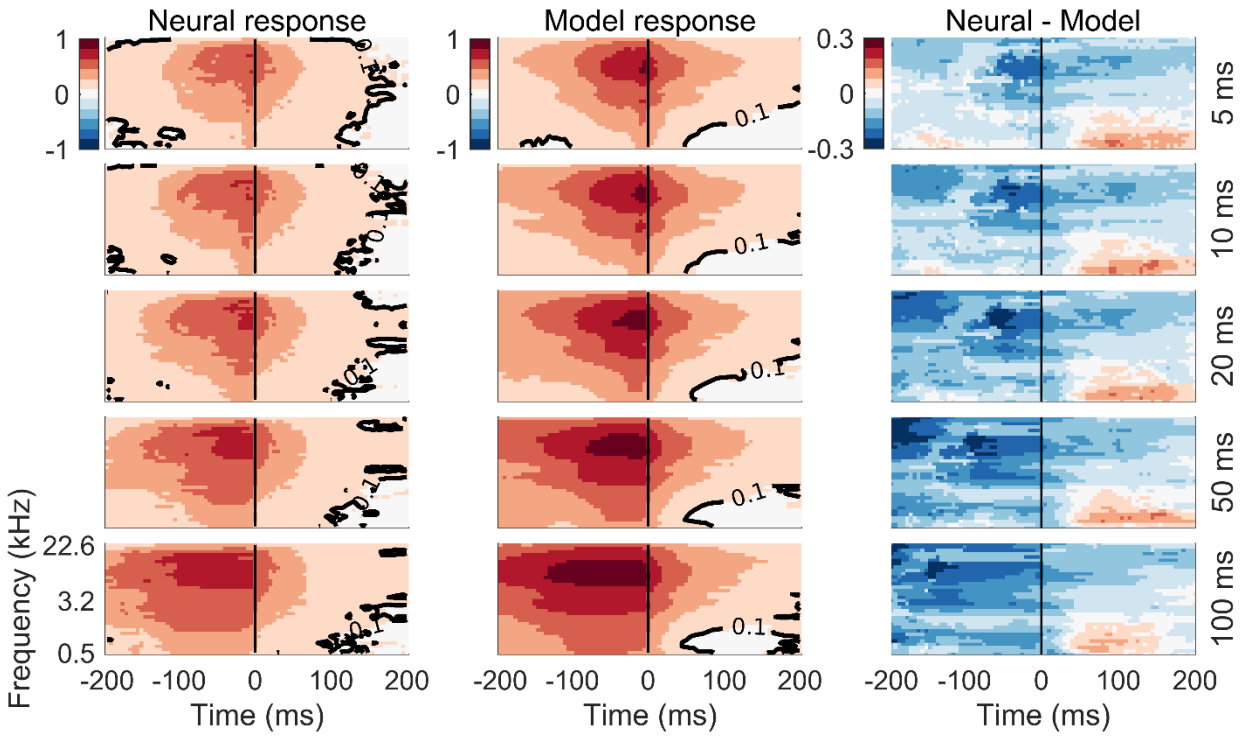

Figure 4.5: Comparison of the reconstruction of past and future stimuli from the population response recorded in A1 to an artificial population response generated by LNP model neurons. *The left column shows the accuracy of stimulus estimation (CC values) from the neural response, the middle column shows the accuracy of stimulus estimation (CC values) from the artificial response and the right column shows the difference between these two. Different rows show the values obtained with different decoding window durations. Colored regions in each plot indicate individual CC values and the black line represents the contour of the plot at a CC value of 0.1.*

4.2.4 Sound frequency selectivity of neurons in the dataset

Our neural dataset represents only a small sample of auditory cortical neurons. Some of the results observed so far might be related to the spectro-temporal characteristics of the neurons in

this dataset. To investigate the characteristics of these neurons, we analyzed the linear STRF of each neuron. We took the linear filters (i.e. the linear STRF) of the LNP encoding models described in the previous section and calculated the mean over all neurons in the population (Fig. 4.6A). This provides the average spectro-temporal characteristics of the neural population. The average STRF (Fig. 4.6A – right panel), of the population shows the typical excitatory field followed by inhibitory field pattern of auditory neurons. The excitatory fields in the STRFs suggest the time-frequency component of the stimuli that excites the neurons the most. We found that the average STRF has excitatory fields spanning the whole range of frequency channels. Furthermore, we also characterized each neuron according to its best frequency (Fig. 4.6B), i.e. the sound frequency component that excites the neuron most. In this case, we identified the best frequency by taking the frequency of the point in the STRF that has maximum excitatory weight. Although the majority of the neurons in our dataset were tuned to high frequencies, low frequency neurons were well represented as well (Fig. 4.6A,B). Thus, neurons in our dataset should be somewhat representative of the A1 frequency selectivity.

We also investigated the level of noise present in the responses of individual neurons since this would also be expected to play a role in the encoding of stimulus information. We used the variance-to-mean spike-count ratio, which is often used as a measure of noise in signal analysis (Fano, 1947), to quantify the noise in the response of each neuron (Eden and Kramer, 2010; Goris et al., 2014). We calculated the mean and variance for the distribution of spike counts (in a 5 ms window) for each neuron. We then plotted a histogram of the ratio of these two quantities for each neuron. This showed that our neurons are mostly super-Poisson in containing noise (Fig. 4.6C). This implies that A1 neurons are more noisy than an LNP model and so an encoder that better emulate the stochastic nature of A1 would be more suitable for our study.

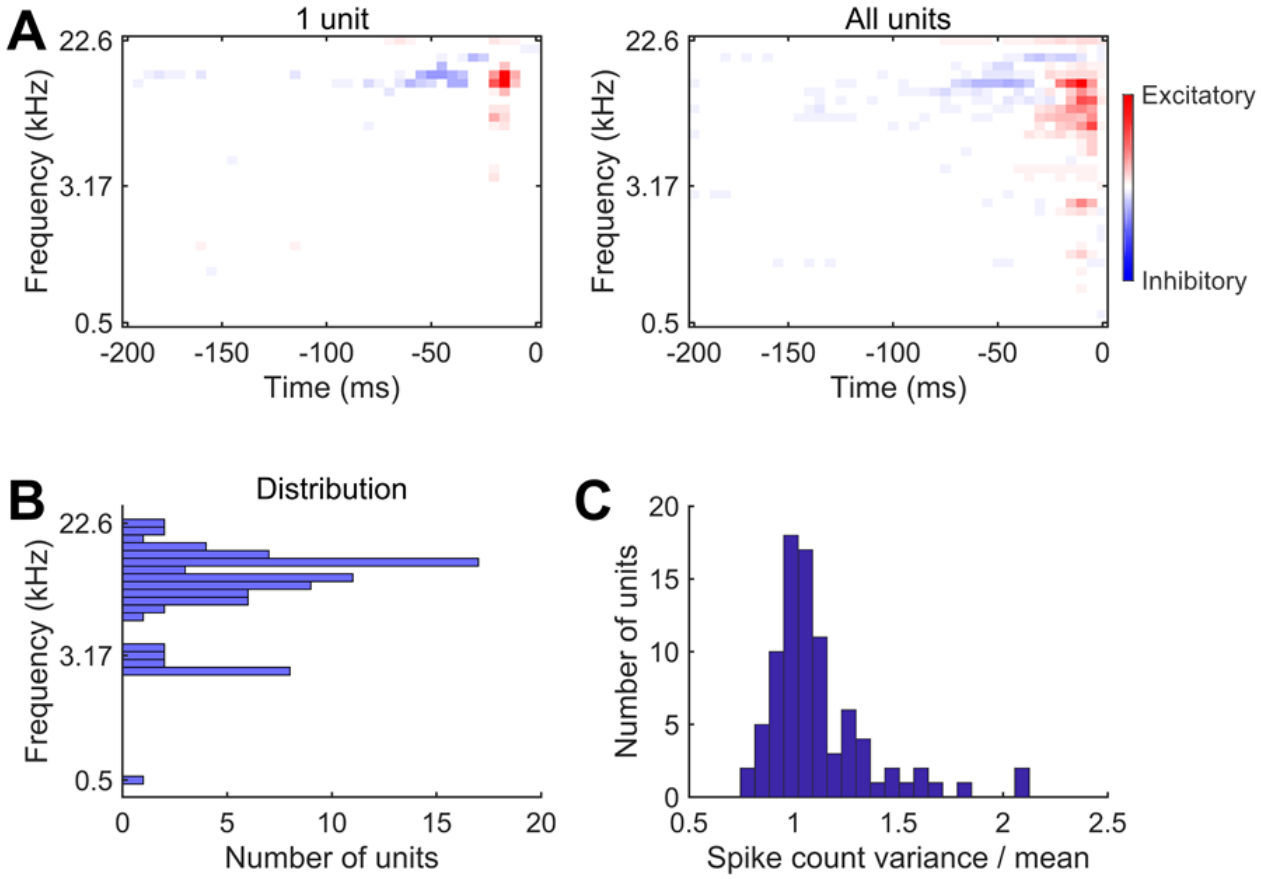

Figure 4.6: Characteristics of neurons in the dataset. *A. The linear STRF of an example neuron (left) and the average STRF of the A1 neural population. The average STRF shows the distribution of spectral and temporal tuning properties of all the neurons in the dataset. B. Distribution of best frequencies of neurons. The best frequency is the frequency field in the STRF where the power is maximum. C. Distribution of the amount of noise present in the responses as measured by the variance/mean of each neuron.*

4.3 Discussion

We showed here that a population of auditory neurons recorded in ferret A1 can estimate several hundred milliseconds of the past (~700 ms) and future (~200 ms) stimulus for a diverse set of natural stimuli. We also showed that the accuracy of stimulus estimation for the past and future improves with the length of the decoding window up to 50-75 ms, after which the decoding accuracy saturates. These findings could help to understand the temporal span of sensory representation in the auditory system. For example, having a brief temporal window of neural response to estimate the future may be useful for guiding rapid actions. Furthermore, we compared the ability of the neural population to estimate the stimulus cochleagram with that of artificial responses produced by an LNP model. Our analysis reveals a superior capacity of the real neural response to predict the future of low frequency components of auditory input compared to the LNP model.

Previous studies have investigated the influence of stimulus history on neural responses in the auditory system. Auditory cortical neurons can contain information about past context for up to few seconds (Asari and Zador, 2009). Attempts have also been made to decode information about past and present stimuli (pure tones) from neural responses (Klampfl et al., 2012). By recording from the auditory cortex while presenting a sequence of pure tones to awake ferrets, Klampfl et al. demonstrated that a linear decoder can estimate both the current and past tones from the neural responses. Moreover, they reported that most information about past and present stimuli is represented in the spiking rate and not in the precise spike timing (Klampfl et al., 2012), which suggests that it would be the most useful to decode the stimulus from mean spiking rates – the approach taken in our study, although precise timing for pure tone response might be less important than it is for natural sounds. Their study provided valuable insights of how neurons

integrate temporal structure of sounds. However, how this relates relate to the capacity of the neural population to predict future inputs has not been explored before.

Prediction is incorporated in various theoretical frameworks that attempt to describe sensory processing, such as, predictive coding (Rao and Ballard, 1999), predictive information (Bialek et al., 2001), slowly varying features (Kayser et al., 2001), and temporal prediction (Singer et al., 2018). Our approach of measuring the encoding of future input is more directly related to predictive information and temporal prediction since these two frameworks posit that neurons in a particular brain area optimally filter in information that is predictive of future input. However, previous study reported that sensory neurons instead encode the surprise or prediction error (Gill et al., 2008). It might be the case that, while majority of A1 neurons encode information to predict the future, some neurons of the population encode prediction error. Identification of prediction error neurons are beyond the scope of our study.

We attempted to show how an individual neuron's temporal dependence on the stimulus history related to the population's capacity to predict the future of the stimulus. If neurons are modelled using an LNP model and the artificial responses produced by the LNP model are used to estimate the stimulus, the estimations by model neurons were generally better across various decoding windows than those of the real neurons. However, we consistently observed a low frequency region of the future cochleagram that seems to be better estimated by the neural population, despite those responses being noisier - with the ratio of the mean to variance of firing rate of neurons being larger than that of the LNP model (mean to variance ratio of Poisson noise is 1). This finding suggests that expectations about the future may be encoded using some nonlinear mechanism that cannot be captured by the LNP model and also perhaps by filtering out some information about the past. This may involve nonlinear encoding schemes previously described

in auditory cortical neurons (Ahrens et al., 2008; Atencio et al., 2008; David and Shamma, 2013; Willmore et al., 2016) or a previously unexplored mechanism.

Furthermore, our results suggest that the representation of predictions of the future might be brief (5-10 ms) at the level of A1. This would make sense if we consider how information about future stimuli or inputs might be used to guide actions. Downstream neurons that depend on a long integration window would respond sluggishly to a sudden change in the stimulus. Any information about future stimuli would facilitate swift action if it is linearly decodable from a brief moment in time. This should improve the accuracy and speed of subsequent behavioral responses.

In more general terms, it is an interesting avenue of analysis to study the relationship between the computation that happens in a single neuron to the computation that happens in the population. Our approach of generating artificial responses using LNP models of individual neurons and then to compare them with real population responses is a useful methodology for studying population coding in a brain area. We show that single neurons' tuning properties and non-linearities inferred from using an LNP model are not sufficient to explain the neural population's ability to predict the future of the stimulus. To account for this, one potential improvement could be to use more complex models of neurons that capture long-term temporal dependence on stimulus history and the complexity of their input due to brainstem processing. However, it could also be that the population's ability to predict the future rests on the connectivity between neurons. A generalized linear model with coupling filters (Pillow et al., 2008) that takes into account the functional connectivity of individual neurons might be a good future direction for an encoding model. Another possibility would be to use deep and recurrent network models of the neural population.

Although our sample of the A1 neurons spanned a large range of best frequencies, adding more low frequency neurons to evenly represent the full neural population would be desirable. Nevertheless, our finding that the recorded neural population predicts better at low frequencies compared to the LNP model may be related to the observation that real neurons can encode sound information from off-best frequency better than an LN model (Harper et al., 2016; Rahman et al., 2019). It is important to note that these results are still preliminary, and more investigation is necessary to understand this phenomenon in detail.

The quality of stimulus estimates may be influenced by the limited set of auditory stimuli in our dataset. Consequently, analysis of a second dataset might improve the generality of our results. Furthermore, a dataset recorded from awake ferrets should also help us understand our findings in a more naturalistic setting. In fact, brain state and brain activity are altered by general anesthesia and we have shown in this thesis that the prediction performance for natural sounds of the same encoding model is different in awake vs anesthetised animals (see Chapter 2). Previous studies suggest that anaesthesia can impact temporal aspects of neural activity in the brain, suggesting that the temporal window of integration of acoustic inputs may be different in awake and anesthetized animals (Lennox, 1949; Sellers et al., 2015). The capacity of auditory neurons to encode temporal expectations and therefore predict the future stimulus may also be compromised, as functional interactions between brain areas are altered in animals under general anesthesia (John et al., 2001; Moeller et al., 2009). A comparison of the decoding results generated using recordings from awake and anesthetized ferrets may serve to advance our understanding on how sensory processing differs under different states of consciousness.

Finally, we could also potentially use a different decoding model, such as a non-linear decoding model. Although we have assumed a linear decoder for the auditory cortical responses, a non-linear decoder might produce more insights. We could also decode sound properties other than

just the power from the cochleagram. For example, we could decode the variance of the cochleagram in a given time window, since prediction of the future might go beyond just estimating the power of the stimulus in a given frequency band and could also be about estimating higher statistics. In fact, various processes exist for adapting to sound statistics (Dean et al., 2005, 2008; Rabinowitz et al., 2011), that could encode such information at a population level and some of this might only be available to a non-linear decoder.

This study makes use of population decoding to investigate the capacity of A1 to predict future inputs. While it is important to understand the neural processes responsible for the encoding of stimulus history, prediction of future inputs may rely on a different or additional set of mechanisms. Our study shows how we can use a combination of encoding and decoding methodology to investigate such processes. Indeed, it has been suggested that sensory systems may be optimized for prediction of future input, rather than representation of past input, only representing the past in as much as it is useful in predicting the future (Bialek et al., 2001; Singer et al., 2018). This study also opens the opportunity for future analysis, such as using a deep-recurrent neural network or a generalized linear model as an encoding model. We also propose decoding beyond estimating the raw values of the sound cochleagram, such as estimating of variance of the values in the cochleagram.

4.4 Methods

4.4.1 The dataset

Most neurons in the dataset used in this study comprised the 73 A1 units described before in Chapter 2 (natural sound dataset 1) and Chapter 3. For the current chapter we have added a further 13 units from one adult ferret that were recorded for the purpose of this project. This experiment followed the same experimental protocol as the previous experiments (see Methods of Chapter 2 and 3) except that the recordings were carried out using a Neuropixels Phase 3a probe. An Ag/AgCl wire was inserted between the dura and the skull as a reference electrode. Data were acquired at a 30 kHz sampling rate using SpikeGLX software (https://github.com/billkarsh/SpikeGLX) and custom Matlab scripts. The recorded signal was processed offline by first digitally highpass filtering at 150 Hz. Common average referencing was performed to remove noise across electrode channels. Spiking activity was then detected and clustered using Kilosort2 software (https://github.com/MouseLand/Kilosort2). Responses from single units were manually curated using Phy (https://github.com/cortex-lab/phy) if they had stereotypical spike shapes with low variance, and their autocorrelation spike histogram showed a clear refractory period. Spikes from a given cluster were often measurable on 4-6 neighbouring electrode channels, facilitating the isolation of single units. Only well isolated units were included in subsequent analyses. In total 76 single units were isolated across 3 penetrations, among which 13 units had a noise ratio < 40 and were included in the subsequent analysis.

Altogether, 20 sound clips were presented containing human speech in different languages, other animal vocalizations (e.g. ferret) and environmental sounds (e.g. wind and water). All clips were 5 s long with a sampling rate of 48,828.125 Hz and with root mean square intensity ranging from 75 to 82 dB SPL. For each sound clip, for each unit, the number of spikes was counted in each 5 ms time bin and averaged over repeats, to provide a response profile $R(t, n)$, where t is time bin

index, and n is the neuron index, for n = 1 to N neurons, where N = 86. The response profiles for neuron *n* is $R_n(t)$.

4.4.2 Training and testing the linear decoder

We trained the linear decoder function, $g_{f,d}(\tau, n)$, so as to minimize the mean squared error (MSE) between the actual stimulus cochleagram $S_f(t-d)$ and the estimated stimulus $\hat{S}_{f,d}(t)$. To prevent overfitting, we use L_1-norm regularization with the hyperparameter, λ. Thus, to obtain the fitted (trained) decoder we apply the process below:

$$\underset{g}{\operatorname{argmin}}\left(\sum_{m=1}^{\mathrm{M}}\sum_{t=1+D}^{\mathrm{T-D}}\left[S_{f,m}(t-d)-\hat{S}_{f,d,m}(t)\right]^2+\lambda\left\|g_{f,d}(\tau,n)\right\|_1\right)$$

Here we add an additional subscript *m*, which indicates clip number and goes from 1 to M. Hence, $S_{f,m}(t-d)$ and $\hat{S}_{f,d,m}(t)$ are respectively, $S_f(t-d)$ and $\hat{S}_{f,d}(t)$ for clip m. Also, *D* is the magnitude of the largest time deviation, d, used.

Thus, after fitting the decoder we obtain a matrix $g_{f,d}(\tau, n)$ over τ and n. This matrix provides the best linear mapping (given our data and regularization) from the neural population response to the cochleagram at d time bins into the past or the future, and in frequency channel f. We perform this fitting for each deviation d and frequency channel f that we investigate.

Due to the limited stimulus set, we used k-fold procedure (k = 5) for evaluation of decoder performance. For each frequency channel f and time deviation d, 4 of the 20 sounds clips and corresponding responses were randomly chosen as a test set, and the remaining 16 clips and responses were used as a training set. Using the training set, 10 decoders each with different

values of the hyperparameter λ were trained (log space from 10^{-3} to 0.5), and the decoder with the lowest mean squared error in estimating the cochleagrams of the test set was chosen. Then the correlation coefficient between the test set and the chosen decoder's estimate was measured. This entire procedure repeated for 5 times for 5 different non-overlapping test sets, providing 5 correlation coefficients. These 5 correlation coefficients were then averaged to provide the correlation coefficient for a particular frequency channel f and time delay d. These averaged correlation coefficients are the ones used throughout the results. As it is the relative value of the correlation coefficient over the different frequency channels and time delays that matters most, rather than the absolute value, and as all conditions are subject to this same method and have similar complexity, we considered it reasonable to set λ using the test set despite possible slight overfitting. Doing this along with a k-fold procedure allowed the use of the full dataset to provide the correlation coefficient, while limiting processing time. However, we think that it would be better in future developments of this study to take a separate validation set from the training set in order set the lambda and prevent overfitting, and then refit using that lambda on the full training set, although this would increase the processing time needed.

4.4.3 The linear-nonlinear-Poisson encoding model

The linear and non-linear stage of the linear-nonlinear-Poission (LNP) encoding model has been described in detail in Chapter 2 and Chapter 3 (see Methods section of those chapters for more detail), being an LN model. As in the other chapters, the two stages of the LN model were fit separately, in each case by minimising the MSE between neural response and model's estimates of neural response, with the linear stage also subject to L_1 regularization as in the other Chapters. For each fold in the k-fold procedure use to train the decoder, and for each neuron, the same training set was first used to train the linear stage of the encoding model for 12 different regularization strengths (log spaced from 10^{-3} to 1). Then, the encoding model with the lowest

mean squared error in estimating the neuron's responses to the test set was chosen. Then the non-linear stage was fit to this chosen model using the same training set. Finally, the LN model's output for the test set was used as input to a Poisson random number generator (https://uk.mathworks.com/help/stats/poissrnd.html) for 20 repeats, as in the neural data, and then averaged over repeats. This whole process was done for each neuron to provide an artificial neural population response.

Then this artificial neural population response was treated just as the real neural data, with the best fitted decoder found in the same way, and then the correlation coefficient measured on the test set in the same way, for each particular frequency channel f and time delay d. This whole process of fitting the LNP models, then their decoders, and then measuring the correlation coefficients was repeated independently for each of the 5 folds in the k-fold procedure used for the real neural responses, providing 5 correlation coefficients. These 5 correlation coefficients were then averaged to provide the correlation coefficient for each frequency channel f and time delay d. These averaged correlation coefficients are the ones used throughout the results.

Chapter 5 | Discussion

Perception of sound by the brain involves extracting information from sound stimuli and transforming their neural representation in ways that are relevant for cognition and behavior. A central aim of auditory neuroscience is to study this transformation by investigating the relationship between sound stimuli and the activity of neurons. In this thesis, we investigated the responses of the auditory cortex to sound stimuli using various encoding and decoding models. We have gathered data from ferret auditory cortex and worked on several large neural datasets available from prior studies. These datasets include ferret A1 responses to natural sounds and artificial stimuli (dynamic random chords).

We centered our research aims around the following questions. (1) How can we model the transformation of information from ear to auditory cortex? (2) What kind of biological mechanisms underlie such transformations? (3) How does the ability of a neural population to encode sound stimuli relate to its capacity to predict the future? We addressed these questions by building encoding models that transform sound stimuli into cortical responses and by building decoding models that reconstruct sound spectrograms from cortical responses. We have several findings that are worth noting at this stage. Firstly, simple models better capture auditory input to cortex. This section of our work identified the most basic algorithm that could explain much of the transformations of the acoustic input on its way to the cortex. We also found that incorporation of certain biological mechanisms, such as the subdivision of auditory nerve fibers on the basis of their rate-level functions, could improve prediction performance of encoding models of A1 responses. We also found that various input and output non-linearities are at interplay in the encoding of sounds from the periphery to the cortex. Secondly, the temporal receptive fields of A1 neurons can be explained by a neural network with dynamic units. Neurons in the auditory

cortex perform temporal integration of stimulus history. Our work explored the duration of stimulus history that is required for a linear spectro-temporal receptive field model. We showed that this history dependence can be explained by short duration receptive fields and various integrative properties present in a biological neuron. Thirdly, we expanded our exploration from encoding of sound stimulus history to the capacity of a population of neurons to predict the future. Our results suggest that some additional non-linearity might be present in neurons to predict the future that cannot be captured by the spectro-temporal feature selectivity and an output non-linearity. This could have behavioral relevance since encoding of predicted future stimuli would help in guiding action.

A challenge faced by the brain as a whole is to encode sensory features from the environment and to decode that information to guide behavior and other responses. In that respect, studying encoding and decoding of sensory stimuli is the natural first step to study the brain. In fact, much of the foundational work in neuroscience has relied on studying the response properties of individual neurons to various types of stimuli (Hartline, 1940; Galambos and Davis, 1943; Hubel and Wiesel, 1962; Margoliash and Konishi, 1985; Passaglia et al., 1997; Gallant et al., 1998; Woolley and Casseday, 2004). While these seminal works have identified different types of neuronal response properties in different brain areas, as well as important features of the information flow in the brain such as hierarchical and parallel processing, in more recent times much interest has been directed towards finding a general model for neuronal responses (Meyer et al., 2017). The strategy of finding a general model relies mostly on rich datasets collected using sophisticated experimental techniques that allow us to record from many neurons at the same time and to record from the same neurons for long periods.

As outlined in Chapter 1, modelling can serve to describe the response properties of neurons, identify mechanisms underlying these responses and help to interpret the properties of neurons based on governing principles. Our use of encoding models has been two fold. Firstly, we have used encoding models to explain neural responses to sound stimuli. We have used a quantitative measure for the accuracy of prediction of the deterministic part of neural signal. Secondly, we have related the response properties of neurons to potential mechanistic functions. In the case of decoding, we have done the reverse. Also, decoding allowed us to explore potential underlying principles relating to prediction by the auditory system.

The search for a model that predicts cortical responses well, and therefore accounts for the structure of their receptive fields, has been a long-standing pursuit of auditory neuroscience. We have shown in the second chapter of this thesis that selecting the right peripheral model can have an important influence on the prediction performance of cortical models. We found that the transformation that auditory stimuli go through in the periphery and the type of inputs that are subsequently sent to the cortex can be relatively well described by a simple model of the auditory periphery, rather than many of the current models that incorporate biological properties of the cochlea. This simple model is comprised of a short time Fourier transform followed by summation of frequency components, resulting in a finite number of logarithmically spaced frequency channels, to which a compressive log non-linearity is then applied. However, our study cannot rule out the possibility that the biological models were a poor representation of the ferret auditory periphery. One way to test whether that is the case or that the system indeed produces a simplified representation of sound, is to study the transformations that happen in different stages of a cochlear model and then to directly compare them with the transformations at different stages of the auditory periphery.

When studying neural response data, an important question to ask is how much of the response variability is due to the stimulus and how much is due to other factors, such as the internal state of the animal and the experimental conditions. In our modelling framework, we used a measure called noise ratio to select units that were somewhat driven by the stimulus and to discard the ones that did not produce a reliable response. Furthermore, our prediction performance measure is based on the accuracy of predicting the part of neural response that is reproduced on average over repeats of the stimulus. By the taking the average of trials that are randomly interspersed with other stimuli, we potentially remove the much of the effects of changes in internal states (e.g. the animal's arousal level or focus of attention) on responses to the stimuli in question. While predicting trial averages has its advantages, predicting trial-by-trial responses would be an interesting avenue for future modelling studies. Indeed, this would be a more suitable approach for the study of the influence of brain states on stimulus encoding.

For studying time-varying signals in neural recordings, an important consideration is the size of the time bins of the neural data. If the neural response is divided into long time bins, we are essentially looking at only the average spiking rate, losing all temporal resolution. Dividing the responses into very small time bins (<1 ms) gives excellent time resolution, but requires processing a lot of data and may not average out forms of noise, such as temporal jitter, potentially interfering with model fitting. By dividing the neural response into moderately small time bins, we attempted to retain as much information in the time-varying signal as possible, while also reducing the number of data points and jitter. We used 4-5 ms time bins for our study, a timescale near those suggested by a study of auditory cortical coding precision (Kayser et al., 2010) and about half of those suggested by other studies (Schnupp et al., 2006; Walker et al., 2008). One could also argue that this is sufficiently small to measure stimulus coding because most auditory neurons will not produce additional spikes in smaller time windows. As a result, our trial averages

can be viewed as spike probability for a given stimulus. However, some neurons in our datasets did diverge from this. But this divergence was observed for few neurons only and to a small degree.

Although neurons are embedded in a network, it is challenging to infer the properties of the underlying network from individual neuronal responses. In recent times, artificial neural networks have become an attractive new avenue to model the underlying network. There is a fair amount of similarity between artificial neural networks and networks of real neurons, in the sense of the basic notion of interconnected integrative neural units with thresholding nonlinear activation functions. However, there is also much dissimilarity between the two. Whereas a single neuron is an elaborate structure with potentially extensive dendritic branching and axonal arborization, an artificial neural network is a pared-down version of real neurons at best. Despite this, training artificial neural networks allows us to ask important question relating to the architecture of real neural networks. It has recently been proposed that it might be beneficial to frame every neuroscience question in the term of training and testing deep neural networks (Richards et al., 2019). Our neural network models can be considered in this context. Harper et al. (2016) previously applied a single hidden-layer neural network to predict auditory cortical responses. Unlike Harper et al. (2016), we have used a sigmoid neural network as opposed to tanh neural network. The tanh nonlinearity can produce negative output values, which are hard to interpret biologically. Although a tanh network can be transformed to an equivalent sigmoid network, this remains hard to interpret for deep or complex networks. Our training and test split was also different from Harper et al. (2016) - whereas they trained models on the first several seconds of a stimulus and predicted the last second, we trained models on the full duration of stimuli and tested them on different stimuli. Despite having an arguably more difficult prediction task, our NRF model performs very similarly to NRF model of Harper et al (2016).

Our approach during this project has been to aim for improved prediction performance of our models but also to be able to relate their inner workings to the relevant biology. That is why in both the second and third chapters we have explored multiple models for the same task, enabling us to compare models instantiating different aspects of the biology. Whereas for the auditory periphery we drew parallels between a biologically detailed model and a more descriptive simple model, for the auditory cortex we did the same by drawing comparisons between a linear nonlinear model, a single hidden layer neural network and a neural network with biologically inspired units.

We can speculate what possible future models could produce better prediction performance than the models we have explored so far. We found in the course of this work that when observed carefully, our models failed to accurately mimic the non-linearity of neural responses in some cases. Specifically, at certain time-points relative to the stimulus presentation, real neurons exhibit non-linearities associated with a very sharp increase in firing rate, whereas the non-linearity in our models often failed to produce such features. In future work, it would therefore be useful to explore potential new non-linearities for encoding models. Furthermore, deep and recurrent neural networks might provide a useful avenue to study in this respect. Convolutional neural networks have been used to model the mechanics of human cochlea (Baby et al., 2021) and to study encoding of speech (Keshishian et al., 2020). Although deep neural networks require a lot of training data and can be hard to interpret, these studies have been able to circumvent these issues by creating simulated data, by exploring biologically plausible hyperparameters and by drawing parallels with a simpler model. Our DNet model has the benefit of interpretability since it models the exponentially decaying dynamics using a simple equation. Modelling this using a deep recurrent neural network will perhaps require more complexity because such dynamics will probably be encoded by the interactions of many units in a deep recurrent neural network.

Whether the brain is optimized to have a faithful representation of the stimulus past or to predict the future stimulus is still an open question. Some initial experimental and theoretical work indicates that the feature selectivity of sensory neurons may be optimized for predicting the future (Palmer et al., 2013; Singer et al., 2018) rather than encoding the past. Our decoding work explored the capacity of A1 populations to predict future input in comparison to their capacity to faithfully encode the past. Although we did not find that that the decoding accuracy in general is better for future than the past (unsurprisingly, as the past is known, unlike the future), we did find that estimates by real neurons of future inputs outperformed that of a linear non-linear encoding model fitted to the same neurons, whereas the reverse was true for past inputs. This raises the intriguing possibility that neurons may possess non-linearities optimized to predict the future. One could speculate that this non-linearity is probably embedded in the neurons' network architecture. In other words, it might be that networks of neurons are optimized to predict the future inputs.

We have focused most of our modelling effort on data collected from anaesthetized animals. While animals under anesthesia provide for a well-control experimental condition for studying sensory processing by removing motor or cognitive interference, anesthesia can interfere with neural processing in the auditory cortex (Zurita et al., 1994; Gaese and Ostwald, 2001) although some studies have suggested the contrary (Rabinowitz et al., 2011; Tischbirek et al., 2019). Our work on predicting responses in awake ferrets suggests that the responses of auditory cortex to natural stimuli might be more non linear (Chapter 2). However, this remains an area of uncertainty, particularly when it comes to modelling the responses of cortical neurons and is something that we hope to explore in future.

Our work provides a possible framework to study encoding in other brain areas and sensory modalities. Based on our findings on the auditory periphery, it may that simple algorithmic expressions that summarize the essential computational transformations may apply to other areas of the nervous system. It would clearly be informative to study visual or olfactory systems using this strategy.

Sensory processing is an integral part of animal behavior. Studying auditory processing in the context of behavior will also provide additional clues as to how neurons in the auditory cortex respond to sound. There have been significant advances in studying task-related encoding of auditory stimuli (Fritz et al., 2003; Nienborg and Cumming, 2009; David, 2018; Elgueda et al., 2019; Saderi et al., 2021). Auditory cortex and midbrain are highly influenced by internal states and behavior (Elgueda et al., 2019; Saderi et al., 2021). Our encoding and decoding modelling framework can be easily extended to include behavioral variables (Park et al., 2014; Runyan et al., 2017) or an animal's internal state (Calhoun et al., 2019), and so future modelling efforts should be focused in this direction.

In summary, we have demonstrated that encoding and decoding models of auditory cortex can provide powerful tools to study various research questions relating to the computations underlying auditory processing in the brain. Much of our effort has been on the relationship between neural activity and the underlying network of neurons and biological mechanisms. We have shown that it is possible, at least to some extent, to unravel from the various complexities of the auditory pathway the necessary elemental algorithms performed by it. We have shown how novel modelling could further our understanding of neurons' receptive fields. Furthermore, we have gained insights into the nature of the information that a population of neurons might be

representing about future sounds that the brain expects to hear. Similar approaches could also be applicable to other sensory modalities and other brain areas.

References

Abeles M, Goldstein MH (1970) Functional architecture in cat primary auditory cortex: columnar organization and organization according to depth. J Neurophysiol 33:172–187.

Aertsen AMHJ, Johannesma PIM (1981a) A comparison of the spectro-temporal sensitivity of auditory neurons to tonal and natural stimuli. Biol Cybern 42:145–156.

Aertsen AMHJ, Johannesma PIM (1981b) The spectro-temporal receptive field. Biol Cybern 42:133–143.

Aertsen AMHJ, Johannesma PIM, Hermes DJ (1980) Spectro-temporal receptive fields of auditory neurons in the grassfrog. Biol Cybern 38:235–248.

Ahrens MB, Linden JF, Sahani M (2008) Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. J Neurosci 28:1929–1942.

Aitkin LM, Webster WR, Veale JL, Crosby DC (1975) Inferior colliculus. I. Comparison of response properties of neurons in central, pericentral, and external nuclei of adult cat. J Neurophysiol 38:1196–1207.

Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2019) Towards reconstructing intelligible speech from the human auditory cortex. Sci Rep 9:1–12.

Akram S, Presacco A, Simon JZ, Shamma SA, Babadi B (2016) Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. Neuroimage 124:906–917.

Alain C, Arnott SR, Hevenor S, Graham S, Grady CL (2001) “What” and “where” in the human auditory system. Proc Natl Acad Sci U S A 98:12301–12306.

Anderson LA, Linden JF (2011) Physiological differences between histologically defined subdivisions in the mouse auditory thalamus. Hear Res 274:48–60.

Asari H, Zador AM (2009) Long-lasting context dependence constrains neural encoding models in rodent auditory cortex. J Neurophysiol 102:2638–2656.

Atencio CA, Sharpee TO, Schreiner CE (2008) Cooperative nonlinearities in auditory cortical neurons. Neuron 58:956–966.

Atencio CA, Sharpee TO, Schreiner CE (2009) Hierarchical computation in the canonical auditory cortical circuit. Proc Natl Acad Sci 106:21894–21899.

Baby D, Van Den Broucke A, Verhulst S (2021) A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. Nat Mach Intell:1–10.

Bandyopadhyay S, Reiss LAJ, Young ED (2007) Receptive field for dorsal cochlear nucleus neurons at multiple sound levels. J Neurophysiol 98:3505–3515.

Bartlett EL, Wang X (2011) Correlation of neural response properties with auditory thalamus subdivisions in the awake marmoset. J Neurophysiol 105:2647–2667.

Bennur S, Tsunada J, Cohen YE, Liu RC (2013) Understanding the neurophysiological basis of auditory abilities for social communication: A perspective on the value of ethological paradigms. Hear Res 305:3–9.

Bialek W, Nemenman I, Tishby N (2001) Predictability, complexity, and learning. Neural Comput 13:2409–2463.

Bialek W, Rieke F, De Ruyter Van Steveninck RR, Warland D (1991) Reading a neural code. Science (80-) 252:1854–1857.

Bizley JK, Nodal FR, Nelken I, King AJ (2005) Functional organization of ferret auditory cortex.

Cereb Cortex 15:1637–1653.

Blackwell JM, Geffen MN (2017) Progress and challenges for understanding the function of cortical microcircuits in auditory processing. Nat Commun 8:2165.

Boer E De (1975) Synthetic whole-nerve action potentials for the cat. J Acoust Soc Am 58:1030–1045.

Bregman AS (1990) Auditory scene analysis: The perceptual organization of sound. The MIT Press.

Bruce IC, Erfani Y, Zilany MSA (2018) A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. Hear Res 360:40–54.

Calabrese A, Schumacher JW, Schneider DM, Paninski L, Woolley SMN (2011) A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. PLoS One 6:e16104.

Calabrese A, Woolley SMN (2015) Coding principles of the canonical cortical microcircuit in the avian brain. Proc Natl Acad Sci U S A 112:3517–3522.

Calford MB (1983) The parcellation of the medial geniculate body of the cat defined by the auditory response properties of single units. J Neurosci 3:2350–2364.

Calhoun AJ, Pillow JW, Murthy M (2019) Unsupervised identification of the internal states that shape natural behavior. Nat Neurosci 22:2040–2049.

Carney LH, Li T, McDonough JM (2015) Speech coding in the brain: Representation of vowel formants by midbrain neurons tuned to sound fluctuations. eNeuro 2:1–12.

Chambers JD, Elgueda D, Fritz JB, Shamma SA, Burkitt AN, Grayden DB (2019) Computational

Neural Modeling of Auditory Cortical Receptive Fields. Front Comput Neurosci 13:28.

Chen C, Cheng M, Ito T, Song S (2018) Neuronal organization in the inferior colliculus revisited with cell-type-dependent monosynaptic tracing. J Neurosci 38:3318–3332.

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887–906.

Chichilnisky EJ (2001) A simple white noise analysis of neuronal light responses. Network 12:199–213.

Christianson BG, Sahani M, Linden JF (2011) Depth-dependent temporal response properties in core auditory cortex. J Neurosci 31:12837–12848.

Christianson GB, Sahani M, Linden JF (2008) The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. J Neurosci 28:446–455.

Colburn HS, Carney LH, Heinz MG (2003) Quantifying the information in auditory-nerve responses for level discrimination. J Assoc Res Otolaryngol 4:294–311.

David S V. (2018) Incorporating behavioral and sensory context into spectro-temporal models of auditory encoding. Hear Res 360:107–123.

David S V, Mesgarani N, Fritz JB, Shamma SA (2009) Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J Neurosci 29:3374–3386.

David S V, Shamma SA (2013) Integration over multiple timescales in primary auditory cortex. J Neurosci 33:19154–19166.

Dayan P, Abbott LF (2001) Theoretical neuroscience: computational and mathematical modeling of neural systems. Massachusetts Institute of Technology Press.

de Boer E (1991) Auditory physics. Physical principles in hearing theory. III. Phys Rep 203:125–231.

De Boer E (1969) Encoding of frequency information in the discharge pattern of auditory nerve fibers. Int J Audiol 8:547–556.

de Melo TM, Bevilacqua MC, Costa OA (2012) Speech perception in cochlear implant users with the HiRes 120 strategy: A systematic review. Braz J Otorhinolaryngol 78:129–133.

Dean I, Harper NS, McAlpine D (2005) Neural population coding of sound level adapts to stimulus statistics. Nat Neurosci 8:1684–1689.

Dean I, Robinson BL, Harper NS, McAlpine D (2008) Rapid neural adaptation to sound level statistics. J Neurosci 28:6430–6438.

deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science (80-) 280:1439–1444.

Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85:1220–1234.

Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89.

Douglas RJ, Martin KAC (2004) Neuronal circuits of the neocortex. Annu Rev Neurosci 27:419–451.

Drew PJ, Abbott LF (2006) Models and properties of power-law adaptation in neural systems. J Neurophysiol:826–833.

Eden UT, Kramer MA (2010) Drawing inferences from Fano factor calculations. J Neurosci

Methods 190:149–152.

Eggermont JJ, Aertsen AMHJ, Johannesma PIM (1983) Prediction of the responses of auditory neurons in the midbrain of the grass frog based on the receptive field. Hear Res 10:191–202.

Ehret G (1978) Stiffness gradient along the basilar membrane as a basis for spatial frequency analysis within the cochlea. J Acoust Soc Am 64:1723–1726.

Elgueda D, Duque D, Radtke-Schuller S, Yin P, David S V., Shamma SA, Fritz JB (2019) State-dependent encoding of sound and behavioral meaning in a tertiary region of the ferret auditory cortex. Nat Neurosci 22:447–459.

Escabí MA, Read HL (2005) Neural Mechanisms for Spectral Analysis in the Auditory Midbrain, Thalamus, and Cortex. Int Rev Neurobiol 70:207–252.

Escabí MA, Schreiner CE (2002) Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. J Neurosci 22:4114–4131.

Espejo ML, Schwartz ZP, David S V. (2019) Spectral tuning of adaptation supports coding of sensory context in auditory cortex. PLoS Comput Biol 15:e1007430.

Evans EF, Wilson JP (1975) Cochlear tuning properties: Concurrent basilar membrane and single nerve fiber measurements. Science (80-) 190:1218–1221.

Fano U (1947) Ionization yield of radiations. II. the fluctuations of the number of ions. Phys Rev 72:26–29.

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models. J Stat Softw 33:1–3.

Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal

receptive fields in primary auditory cortex. Nat Neurosci 6:1216–1223.

Gaese BH, Ostwald J (2001) Anesthesia Changes Frequency Tuning of Neurons in the Rat Primary Auditory Cortex. J Neurophysiol 86:1062–1066.

Galambos R, Davis H (1943) The response of single auditory nerve-fibers to acoustic stimulation. J Neurophysiol 6:39–57.

Gallant JL, Connor CE, Van Essen DC (1998) Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. Neuroreport 9.

Gaucher Q, Panniello M, Ivanov AZ, Dahmen JC, King AJ, Walker KMM (2020) Complexity of frequency receptive fields predicts tonotopic variability across species. Elife 9:1–25.

Gielen CCAM, Hesselmans GHFM, Johannesma PIM (1988) Sensory interpretation of neural activity patterns. Math Biosci 88:15–35.

Gilad A, Maor I, Mizrahi A (2020) Learning-related population dynamics in the auditory thalamus. Elife 9:1–18.

Gill P, Woolley SMN, Fremouw T, Theunissen FE (2008) What's that sound? Auditory area CLM encodes stimulus surprise, not intensity or intensity changes. J Neurophysiol 99:2809–2820.

Gill P, Zhang J, Woolley SMN, Fremouw T, Theunissen FE (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21:5–20.

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proc. AISTATS, pp 249–256.

Goris RLT, Movshon JA, Simoncelli EP (2014) Partitioning neuronal variability. Nat Neurosci 17:858–865.

Gourévitch B, Noreña A, Shaw G, Eggermont JJ (2009) Spectrotemporal receptive fields in

anesthetized cat primary auditory cortex are context dependent. Cereb Cortex 19:1448–1461.

Green T, Faulkner A, Rosen S, Macherey O (2005) Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification. J Acoust Soc Am 118:375–385.

Greenberg S, Ainsworth WA (2006) Speech processing in the auditory system: An overview. In: Speech Processing in the Auditory System, pp 1–62. Springer-Verlag.

Griffiths TD, Warren JD (2004) What is an auditory object? Nat Rev Neurosci 5:887–892.

Harper NS, Schoppe O, Willmore BDB, Cui ZF, Schnupp JWH, King AJ (2016) Network receptive field modeling reveals extensive integration and multi-feature selectivity in auditory cortical neurons. PLoS Comput Biol 12:e1005113.

Hartline HK (1940) The receptive fields of optic nerver fibers. Am J Physiol 130:690–699.

Hauser MD, Chomsky N, Fitch WT (2002) The faculty of language: What is it, who has it, and how did it evolve? Science (80-) 298:1569–1579.

Hesselmans GHFM, Johannesma PIM (1989) Spectro-temporal interpretation of activity patterns of auditory neurons. Math Biosci 93:31–51.

Homma NY, Happel MFK, Nodal FR, Ohl FW, King AJ, Bajo VM (2017) A Role for Auditory Corticothalamic Feedback in the Perception of Complex Sounds. J Neurosci 37:6149–6161.

Hsu A, Borst A, Theunissen F (2004) Quantifying variability in neural responses and its application for the validation of model predictions. Netw Comput Neural Syst 15:91–109.

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154.

Hudspeth AJ, Jessell TM, Kandel ER, Schwartz JH, Siegelbaum SA (2013) Principles of neural science.

Islam MA, Jassim WA, Cheok NS, Zilany MSA (2016) A Robust Speaker Identification System Using the Responses from a Model of the Auditory Periphery Zeng F-G, ed. PLoS One 11:e0158520.

Ito T, Bishop DC, Oliver DL (2016) Functional organization of the local circuit in the inferior colliculus. Anat Sci Int 91:22–34.

John ER, Prichep LS, Kox W, Valdés-Sosa P, Bosch-Bayard J, Aubert E, Tom M, DiMichele F, Gugino LD (2001) Invariant reversible QEEG effects of anesthetics. Conscious Cogn 10:165–183.

Johnson KO (2000) Neural coding. Neuron 26:563–566.

Joris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude-modulated sounds. Physiol Rev 84:541–577.

Kadia SC, Wang X (2003) Spectral integration in A1 of awake primates: Neurons with single- and multipeaked tuning characteristics. J Neurophysiol 89:1603–1622.

Kanold PO, Nelken I, Polley DB (2014) Local versus global scales of organization in auditory cortex. Trends Neurosci 37:502–510.

Kates J, Arehart K (2014) The Hearing-Aid Speech Quality Index (HASQI) Version 2. J Audio Eng Soc 62:99–117.

Kayser C, Einhäuser W, Dümmer O, König P, Körding K (2001) Extracting slow subspaces from natural videos leads to complex cells. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp 1075–1080. Springer Verlag.

Kayser C, Logothetis NK, Panzeri S (2010) Millisecond encoding precision of auditory cortex neurons. Proc Natl Acad Sci U S A 107:16976–16981.

Keshishian M, Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2020) Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. Elife 9:1–24.

Klampfl S, David S V., Yin P, Shamma SA, Maass W (2012) A quantitative analysis of information about past and present stimuli encoded by spikes of A1 neurons. J Neurophysiol 108:1366–1380.

Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J Comput Neurosci 9:85–111.

Kozlov AS, Gentner TQ (2016) Central auditory neurons have composite receptive fields. Proc Natl Acad Sci 113:1441–1446.

Laudanski J, Edeline J-M, Huetz C (2012) Differences between spectro-temporal receptive fields eerived from artificial and natural stimuli in the auditory cortex Malmierca MS, ed. PLoS One 7:e50539.

Lauteslager T, O'Sullivan JA, Reilly RB, Lalor EC (2014) Decoding of attentional selection in a cocktail party environment from single-trial EEG is robust to task. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014, pp 1318–1321. Institute of Electrical and Electronics Engineers Inc.

Lennox WG (1949) Influence of drugs on the human electroencephalogram. Electroencephalogr Clin Neurophysiol 1:45–51.

Lim HH, Lenarz M, Lenarz T (2009) Auditory Midbrain Implant: A Review. Trends Amplif 13:149–

180.

Linden JF (2003) Columnar transformations in auditory Cortex? A comparison to visual and somatosensory cortices. Cereb Cortex 13:83–89.

Linden JF, Liu RC, Sahani M, Schreiner CE, Merzenich MM (2003) Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. J Neurophysiol 90:2660–2675.

Lyon RF (1982) A computational model of filtering, detection, and compression in the cochlea. In: ICASSP ’82. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp 1282–1285. Institute of Electrical and Electronics Engineers.

Lyon RF (2011) Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function. J Acoust Soc Am 130:3893–3904.

Machens CK, Wehr MS, Zador AM (2004) Linearity of cortical receptive fields measured with natural sounds. J Neurosci 24:1089–1100.

Maier JK, Hehrmann P, Harper NS, Klump GM, Pressnitzer D, McAlpine D (2012) Adaptive coding is constrained to midline locations in a spatial listening task. J Neurophysiol 108:1856–1868.

Margoliash D, Fortune ES (1992) Temporal and harmonic combination-sensitive neurons in the zebra finch’s HVc. J Neurosci 12:4309–4326.

Margoliash D, Konishi M (1985) Auditory representation of autogenous song in the song system of white-crowned sparrows. Proc Natl Acad Sci U S A 82:5997–6000.

Marr D (1982) Vision: A Computational Investigation into the human representation and processing of visual information. New York, NY WH Free Co.

Marr D, Poggio T (1976) From understanding computation to understanding neural circuitry.

Artificial Intelligence Laboratory. A.I. Memo. Massachusetts Institute of Technology.

Martin S, Brunner P, Holdgraf C, Heinze H-J, Crone NE, Rieger J, Schalk G, Knight RT, Pasley BN (2014) Decoding spectrotemporal features of overt and covert speech from the human cortex. Front Neuroeng 7:14.

Massoudi R, Van Wanrooij MM, Versnel H, Van Opstal AJ (2015) Spectrotemporal response properties of core auditory cortex neurons in awake monkey Malmierca MS, ed. PLoS One 10:e0116118.

McDermott JH, Simoncelli EP (2011) Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron 71:926–940.

McFarland JM, Cui Y, Butts DA (2013) Inferring nonlinear neuronal computation based on physiologically plausible inputs Bethge M, ed. PLoS Comput Biol 9:e1003143.

Meddis R (2006) Auditory-nerve first-spike latency and auditory absolute threshold: A computer model. J Acoust Soc Am 119:406–417.

Meddis R, Lecluyse W, Clark NR, Jürgens T, Tan CM, Panda MR, Brown GJ (2013) A computer model of the auditory periphery and its application to the study of hearing. In: Basic Aspects of Hearing (Moore BCJ, Patterson RD, Winter IM, Carlyon RP, Gockel HE, eds), pp 11–20. Springer, New York, NY.

Meddis R, O'Mard LP, Lopez-Poveda EA (2001) A computational algorithm for computing nonlinear auditory frequency selectivity. J Acoust Soc Am 109:2852–2861.

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236.

Mesgarani N, David S V., Fritz JB, Shamma SA (2014) Mechanisms of noise robust representation of speech in primary auditory cortex. Proc Natl Acad Sci U S A 111:6792–

6797.

Mesgarani N, David S V, Fritz JB, Shamma SA (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J Neurophysiol 102:3329–3339.

Meyer AF, Williamson RS, Linden JF, Sahani M (2017) Models of neuronal stimulus-response functions : elaboration, estimation, and evaluation. Front Syst Neurosci 10:109.

Miller LM, Escabí MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87:516–527.

Mitani A, Shimokouchi M, Itoh K, Nomura S, Kudo M, Mizuno N (1985) Morphology and laminar organization of electrophysiologically identified neurons in the primary auditory cortex in the cat. J Comp Neurol 235:430–447.

Moeller S, Nallasamy N, Tsao DY, Freiwald WA (2009) Functional connectivity of the macaque brain across stimulus and arousal states. J Neurosci 29:5897–5909.

Narayan SS, Temchin AN, Recio A, Ruggero MA (1998) Frequency Tuning of Basilar Membrane and Auditory Nerve Fibers in the Same Cochleae. Science (80-) 282:1882–1884.

Nelken I, Bizley J, Shamma SA, Shamma SA, Wang XQ, Wang XQ (2014) Auditory cortical processing in real-world listening: the auditory system going real. J Neurosci 34:15135–15138.

Nelken I, Rotman Y, Yosef OB (1999) Responses of auditory-cortex neurons to structural features of natural sounds. Nature 397:154–157.

Nie K, Barco A, Zeng F-G (2006) Spectral and temporal cues in cochlear implant speech perception. Ear Hear 27:208–217.

Nienborg H, Cumming BG (2009) Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature 459:89–92.

Nilsson HG, Møller AR (1977) Linear and nonlinear models of the basilar membrane motion. Biol Cybern 27:107–112.

Nodal FR, King AJ (2014) Hearing and auditory function in ferrets. In: Biology and Diseases of the Ferret, pp 685–710. Ames, USA: John Wiley & Sons, Inc.

O'Sullivan J, Chen Z, Herrero J, McKhann GM, Sheth SA, Mehta AD, Mesgarani N (2017) Neural decoding of attentional selection in multi-speaker environments without access to clean sources. J Neural Eng 14:056001.

Oonishi S, Katushi Y (1965) Functional organization and integrative mechanism of the auditory cortex of the cat. Jpn J Physiol 15:342–365.

Osen KK (1969) Cytoarchitecture of the cochlear nuclei in the cat. J Comp Neurol 136:453–483.

Palmer SE, Marre O, Berry MJ, Bialek W (2013) Predictive information in a sensory population. Proc Natl Acad Sci 112:6908–6913.

Paninski L (2004) Maximum likelihood estimation of cascade point-process neural encoding models. Netw Comput Neural Syst 15:243–262.

Park IM, Meister MLR, Huk AC, Pillow JW (2014) Encoding and decoding in parietal cortex during sensorimotor decision-making. Nat Neurosci 17:1395–1403.

Pasley BN, David S V., Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex Zatorre R, ed. PLoS Biol 10:e1001251.

Passaglia C, Dodge F, Herzog E, Jackson S, Barlow R (1997) Deciphering a neural code for

vision. Proc Natl Acad Sci U S A 94:12649–12654.

Peruzzi D, Sivaramakrishnan S, Oliver DL (2000) Identification of cell types in brain slices of the inferior colliculus. Neuroscience 101:403–416.

Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, Simoncelli EP (2008) Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature 454:995–999.

Prather JF (2013) Auditory signal processing in communication: perception and performance of vocal sounds. Hear Res 305:144–155.

Puvvada KC, Simon JZ (2017) Cortical representations of speech in a multitalker auditory scene. J Neurosci 37:9189–9196.

Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH (2013) Constructing noise-invariant representations of sound in the auditory pathway.

Rabinowitz NC, Willmore BDB, Schnupp JWH, King AJ (2011) Contrast gain control in auditory cortex. Neuron 70:1178–1191.

Rabinowitz NC, Willmore BDB, Schnupp JWH, King AJ (2012) Spectrotemporal contrast kernels for neurons in primary auditory cortex. J Neurosci 32:11271–11284.

Rahman M, Willmore BDB, King AJ, Harper NS (2019) A dynamic network model of temporal receptive fields in primary auditory cortex. PLOS Comput Biol 15:e1006618.

Ramirez AD, Ahmadian Y, Schumacher J, Schneider D, Woolley SMN, Paninski L (2011) Incorporating naturalistic correlation structure improves spectrogram reconstruction from neuronal activity in the songbird auditory midbrain. J Neurosci 31:3828–3842.

Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of

some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Read HL, Winer JA, Schreiner CE (2002) Functional architecture of auditory cortex. Curr Opin Neurobiol 12:433–440.

Real E, Asari H, Gollisch T, Meister M (2017) Neural circuit inference from function to structure. Curr Biol 27:189–198.

Recanzone GH, Schreiner CE, Sutter ML, Beitel RE, Merzenich MM (1999) Functional organization of spectral receptive fields in the primary auditory cortex of the owl monkey. J Comp Neurol 415:460–481.

Recio A, Rhode WS (2000) Basilar membrane responses to broadband stimuli. J Acoust Soc Am 108:2281–2298.

Recio A, Rich NC, Narayan SS, Ruggero MA (1998) Basilar-membrane responses to clicks at the base of the chinchilla cochlea. J Acoust Soc Am 103:1972–1989.

Reid RC, Soodak RE, Shapley RM (1987) Linear mechanisms of directional selectivity in simple cells of cat striate cortex. Proc Natl Acad Sci 84:8740–8744.

Rhode WS (1971) Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys using the Mössbauer Technique. J Acoust Soc Am 49:1218–1231.

Rhode WS, Smith PH (1986) Encoding timing and intensity in the ventral cochlear nucleus of the cat. J Neurophysiol 56:261–286.

Richards BA et al. (2019) A deep learning framework for neuroscience. Nat Neurosci 22:1761–1770.

Robinson BL, Harper NS, McAlpine D (2016) Meta-adaptation in the auditory midbrain under cortical influence. Nat Commun 7:1–8.

Robles L, Ruggero MA (2001) Mechanics of the mammalian cochlea. Physiol Rev 81:1305–1352.

Ru P (2001) Multiscal multirate spectro-temporal auditory model.

Rubio ME (2018) Microcircuits of the Ventral Cochlear Nucleus. In, pp 41–71. Springer Handbook of Auditory Research.

Ruggero MA (1992) Responses to sound of the basilar membrane of the mammalian cochlea. Curr Opin Neurobiol 2:449–456.

Ruggero MA, Rich NC (1986) Basilar membrane mechanics at the base of the chinchilla cochlea. I. Input-output functions, tuning curves, and response phasesa). J Acoust Soc Am 80:1364–1374.

Runyan CA, Piasini E, Panzeri S, Harvey CD (2017) Distinct timescales of population coding across cortex. Nature 548:92–96.

Russo M, Stella M, Sikora M, Pekić V (2019) Robust cochlear-model-based speech recognition. Computers 8.

Ryugo DK, Menotti-Raymond M (2012) Feline Deafness. Vet Clin North Am Small Anim Pract 42:1179–1207.

Sachs MB, Abbas PJ (1978) Rate versus level functions for auditory-nerve fibers in cats: Bandlimited noise bursts. J Acoust Soc Am 64:S135–S135.

Sadagopan S, Wang X (2009) Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. J Neurosci 29:11192–11202.

Saderi D, Schwartz ZP, Heller CR, Pennington JR, David S V (2021) Dissociation of task engagement and arousal effects in auditory cortex and midbrain. Elife 10.

Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014)

Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput Biol 10.

Saremi A, Stenfelt S (2013) Effect of metabolic presbyacusis on cochlear responses: A simulation approach using a physiologically-based model. J Acoust Soc Am 134:2833–2851.

Schlauch RS, DiGiovanni JJ, Ries DT (1998) Basilar membrane nonlinearity and loudness. J Acoust Soc Am 103:2010–2020.

Schnupp J, Nelken I, King A (2011) Auditory neuroscience : making sense of sound. MIT Press.

Schnupp JWH, Hall TM, Kokelaar RF, Ahmed B (2006) Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. J Neurosci 26:4785–4795.

Schnupp JWH, Mrsic-Flogel TD, King AJ (2001) Linear processing of spatial cues in primary auditory cortex. Nature 414:200–204.

Schofield BR, Beebe NL (2019) Subtypes of GABAergic cells in the inferior colliculus. Hear Res 376:1–10.

Schoppe O, Harper NS, Willmore BDB, King AJ, Schnupp JWH (2016) Measuring the performance of neural models. Front Comput Neurosci 10:1–11.

Schreiner CE, Winer JA (2007) Auditory cortex mapmaking: principles, projections, and plasticity. Neuron 56:356–365.

Sellers KK, Bennett D V., Hutt A, Williams JH, Fröhlich F (2015) Awake vs. anesthetized: layer-specific sensory processing in visual cortex and functional connectivity between cortical areas. J Neurophysiol 113:3798–3815.

Shamma SA, Fleshman JW, Wiser PR, Versnel H (1993) Organization of response areas in ferret primary auditory cortex. J Neurophysiol 69:367–383.

Sharpee T, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput 16:223–250.

Sharpee TO (2013) Computational identification of receptive fields. Annu Rev Neurosci 36:103–120.

Sharpee TO, Atencio CA, Schreiner CE (2011) Hierarchical representations in the auditory cortex. Curr Opin Neurobiol 21:761–767.

Simoncelli EP, Paninski L, Pillow J, Schwartz O (2004) Characterization of neural responses with stochastic stimuli. In: The New Cognitive Neurosciences, 3rd ed. (Gazzaniga M, ed), pp 327–338. MIT Press.

Singer Y, Teramoto Y, Willmore BDB, King AJ, Schnupp JWH, Harper NS (2018) Sensory cortex is optimised for prediction of future input. Elife 7:1–31.

Sohl-Dickstein J, Poole B, Ganguli S (2013) Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods. arXiv Prepr arXiv13112115.

Steadman MA, Sumner CJ (2018) Changes in neuronal representations of consonants in the ascending auditory system and their role in speech recognition. Front Neurosci 12:671.

Stevenson IH, London BM, Oby ER, Sachs NA, Reimer J, Englitz B, David S V., Shamma SA, Blanche TJ, Mizuseki K, Zandvakili A, Hatsopoulos NG, Miller LE, Kording KP (2012) Functional connectivity and tuning curves in populations of simultaneously recorded neurons. PLoS Comput Biol 8:1002775.

Suga N, O'Neil WE, Manabe T (1978) Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the mustache Bat. Science (80-) 200:778–782.

Sumner CJ, Lopez-Poveda EA, O'Mard LP, Meddis R (2002) A revised model of the inner-hair cell and auditory-nerve complex. J Acoust Soc Am 111:2178.

Sumner CJ, O'Mard LP, Lopez-poveda EA, Meddis R (2003) A nonlinear filter-bank model of the guinea-pig cochlear nerve: Rate responses. J Acoust Soc Am 113:3264.

Sumner CJ, Palmer AR (2012) Auditory nerve fibre responses in the ferret. Eur J Neurosci 36:2428–2439.

Sumner CJ, Wells TT, Bergevin C, Sollini J, Kreft HA, Palmer AR, Oxenham AJ, Shera CA (2018) Mammalian behavior and physiology converge to confirm sharper cochlear tuning in humans. Proc Natl Acad Sci U S A 115:11322–11326.

Sutter ML, Schreiner CE (1991) Physiology and topography of neurons with multipeaked tuning curves in cat primary auditory cortex. J Neurophysiol 65:1207–1226.

Tan X, Young H, Matic AI, Zirkle W, Rajguru S, Richter C-P (2015) Temporal properties of inferior colliculus neurons to photonic stimulation in the cochlea. Physiol Rep 3:e12491.

Theunissen E, David S V, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural. Netw Comput Neural Syst 12:289–316.

Theunissen E, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons. J Neurosci 20:2315–2331.

Theunissen FE, Elie JE (2014) Neural processing of natural sounds. Nat Rev Neurosci 15:355–366.

Thompson J (2013) Neural decoding of subjective music listening experiences.

Thorpe WH (1972) The comparison of vocal communication in animals and man. In: Non-verbal communication (Hinde RA, ed), pp 27–48. Cambridge University Press.

Thorson IL, Liénard J, David S V (2015) The essential complexity of auditory receptive fields.

PLoS Comput Biol 11:e1004628.

Tischbirek CH, Noda T, Tohmi M, Birkner A, Nelken I, Konnerth A (2019) In vivo functional mapping of a cortical column at single-neuron resolution. Cell Rep 27:1319-1326.e5.

Trussell LO, Oertel D (2018) Microcircuits of the Dorsal Cochlear Nucleus. In, pp 73–99. Springer Handbook of Auditory Research.

Vasquez-Lopez SA, Weissenberger Y, Lohse M, Keating P, King AJ, Dahmen JC (2017) Thalamic input to auditory cortex is locally heterogeneous but globally tonotopic. Elife 6.

Verhulst S, Altoè A, Vasilkov V (2018) Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss. Hear Res 360:55–75.

von Békésy G, Peake WT (1990) Experiments in hearing. J Acoust Soc Am 88:2905–2905.

Walker KMM, Ahmed B, Schnupp JWH (2008) Linking cortical spike pattern codes to auditory perception. J Cogn Neurosci 20:135–152.

Wallace MN, Coomber B, Sumner CJ, Grimsley JMS, Shackleton TM, Palmer AR (2011) Location of cells giving phase-locked responses to pure tones in the primary auditory cortex. Hear Res 274:142–151.

Wallace MN, Palmer AR (2008) Laminar differences in the response properties of cells in the primary auditory cortex. Exp Brain Res 184:179–191.

Wallace MN, Rutkowski RG, Shackleton TM, Palmer AR (2000) Phase-locked responses to pure tones in guinea pig auditory cortex. Neuroreport 11:3989–3993.

Wallace MN, Shackleton TM, Palmer AR (2002) Phase-locked responses to pure tones in the primary auditory cortex. Hear Res 172:160–171.

Wang K, Shamma S (1994) Self-normalization and noise-robustness in early auditory

representations. IEEE Trans Speech Audio Process 2:421–435.

Wang K, Shamma SA (1995) Auditory analysis of spectro-temporal information in acoustic signals. IEEE Eng Med Biol Mag 14:186–194.

Wen B, Wang GI, Dean I, Delgutte B (2009) Dynamic range adaptation to sound level statistics in the auditory nerve. J Neurosci 29:13797–13808.

Williams RJ, Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1:270–280.

Williamson RS, Ahrens MB, Linden JF, Sahani M (2016) Input-specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds. Neuron 91:467–481.

Willmore BDB, Schoppe O, King AJ, Schnupp JWH, Harper NS (2016) Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. J Neurosci 36:280–289.

Winer JA, Miller LM, Lee CC, Schreiner CE (2005) Auditory thalamocortical transformation: Structure and function. Trends Neurosci 28:255–263.

Winter IM, Robertson D, Yates GK (1990) Diversity of characteristic frequency rate-intensity functions in guinea pig auditory nerve fibres. Hear Res 45:191–202.

Wirtzfeld MR, Pourmand N, Parsa V, Bruce IC (2017) Predicting the quality of enhanced wideband speech with a cochlear model. J Acoust Soc Am 142:EL319.

Wong DDE, Fuglsang SA, Hjortkjær J, Ceolini E, Slaney M, de Cheveigné A (2018) A comparison of regularization methods in forward and backward models for auditory attention decoding. Front Neurosci 12:531.

Woolley SMN, Casseday JH (2004) Response properties of single neurons in the zebra finch auditory midbrain: response patterns, frequency coding, intensity coding, and spike latencies. J Neurophysiol 91:136–151.

Yates GK, Winter IM, Robertson D (1990) Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range. Hear Res 45:203–219.

Yildiz IB, Mesgarani N, Deneve S (2016) Predictive ensemble decoding of acoustical features explains context-dependent receptive fields. J Neurosci 36:12338–12350.

Zeng FG, Rebscher S, Harrison W, Sun X, Feng H (2008) Cochlear implants: System design, integration, and evaluation. IEEE Rev Biomed Eng 1:115–142.

Zilany MSA, Bruce IC, Carney LH (2014) Updated parameters and expanded simulation options for a model of the auditory periphery. J Acoust Soc Am 135:283–286.

Zilany MSA, Bruce IC, Nelson PC, Carney LH (2009) A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. J Acoust Soc Am 126:2390–2412.

Zilany MSA, Carney LH (2010) Power-law dynamics in an auditory-nerve model can account for neural adaptation to sound-level statistics number of occurrences. J Neurosci 30:10380–10390.

Zurita P, Villa AEP, de Ribaupierre Y, de Ribaupierre F, Rouiller EM (1994) Changes of single unit activity in the cat's auditory thalamus and cortex associated to different anesthetic conditions. Neurosci Res 19:303–316.