COMPUTATIONAL MODELS OF REPRESENTATION AND PLASTICITY IN THE CENTRAL AUDITORY SYSTEM

by

Michael A. Carlin

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

February, 2015

Abstract

The performance for automated speech processing tasks like speech recognition and speech activity detection rapidly degrades in challenging acoustic conditions. It is therefore necessary to engineer systems that extract meaningful information from sound while exhibiting invariance to background noise, different speakers, and other disruptive channel conditions. In this thesis, we take a biomimetic approach to these problems, and explore computational strategies used by the central auditory system that underlie neural information extraction from sound.

In the first part of this thesis, we explore coding strategies employed by the central auditory system that yield neural responses that exhibit desirable noise robustness. We specifically demonstrate that a coding strategy based on sustained neural firings yields richly structured spectro-temporal receptive fields (STRFs) that reflect the structure and diversity of natural sounds. The emergent receptive fields are comparable to known physiological neuronal properties and can be employed as a signal processing strategy to improve noise invariance in a speech recognition task.

Next, we extend the model of sound encoding based on spectro-temporal receptive fields to incorporate the cognitive effects of selective attention. We propose a framework for modeling attention-driven plasticity that induces changes to receptive fields driven by task demands. We define a discriminative cost function whose optimization and solution reflect a biologically plausible strategy for STRF adaptation that helps listeners better attend to target sounds. Importantly, the adaptation patterns predicted by the framework have a close correspondence with known neurophysiological

ABSTRACT

data. We next generalize the framework to act on the spectro-temporal dynamics of task-relevant stimuli, and make predictions for tasks that have yet to be experimentally measured. We argue that our generalization represents a form of object-based attention, which helps shed light on the current debate about auditory attentional mechanisms. Finally, we show how attention-modulated STRFs form a high-fidelity representation of the attended target, and we apply our results to obtain improvements in a speech activity detection task.

Overall, the results of this thesis improve our general understanding of central auditory processing, and our computational frameworks can be used to guide further studies in animal models. Furthermore, our models inspire signal processing strategies that are useful for automated speech and sound processing tasks.

Primary Reader: Dr. Mounya Elhilali

Secondary Reader: Dr. Andreas Andreou

Acknowledgments

This experience would not have been possible without the help from many people. To Mounya, thank you for your guidance, mentorship, patience, and for giving me the freedom to explore many ideas over the years. To Andreas and Ralph, thank you for sitting on my dissertation committee and providing your expertise. Many thanks are also due to Shihab Shamma, Jonathan Fritz, and the Neural Systems Laboratory at the University of Maryland, College Park for providing data used in much of the analysis presented in this thesis. Lastly, I'm grateful for financial support from a fellowship from the Human Language Technology Center of Excellence, and grants from the National Science Foundation, National Institutes of Health, Air Force Office of Scientific Research, and Office of Naval Research.

To my friends in the Laboratory for Computational Audio Perception: Mike, Susan, Sridhar, Kailash, Merve, Dimitra, Debmalya, Nick, Ashwin, and Ben. You were all a pleasure to work with, and thank you for letting me fill up all the whiteboards. To my friends and colleagues in the Center for Language and Speech Processing, especially: Anni, Keith, Jonny, Ming, Sivaram, Sam, Scott, Carolina, and Ruth. Thanks for making CLSP a great place to work and be challenged. To my friends in the ECE department, especially: Chris, John, Keith, Kevin, Andrew, Tom, and Dan. Thanks for the pizza parties, softball (TSRP!), beer tastings, coffee, lifting of heavy things, and other shenanigans over the years. To my colleagues at the Human Language Technology Center of Excellence: Aren, Alan, Peggy, Ben, and the many friends met through the SCALE workshops. Thank you for helping me be mindful of how my work can have impact on important problems.

ACKNOWLEDGMENTS

Thanks are also due to many mentors over the years, especially Drs. Robert Yantorno and Stanley Wenndt. Thank you for exposing me to the world of speech and audio, and for convincing me that pursuing a Ph.D. was the right thing to do.

To Mom, Dad, Jay, Matt, and Lindsay, thank you for all your love and support, and for sticking this out with me over the years; I promise I'm finished with school.

Finally, none of this would have been possible without the love and support of my family. To Amanda, there are no words to fully express my gratitude for everything you've done and have been for me throughout this experience — thank you from the bottom of my heart. And to Patrick, you've changed my life, put everything in perspective, and are my greatest joy. Let me know if and when you ever read this.

Dedication

This work is dedicated to Amanda and Patrick, the most important people in my life.

Contents

List of Tables

List of Figures

LIST OF FIGURES

Chapter 1

Introduction

Sound conveys much about the world around us, carrying acoustic information that we can experience as the pleasure of music or conversation with a friend. The auditory system provides us access to sound, and is remarkable in its ability to make sense of complex acoustic scenes. For example, human listeners effortlessly recognize speech regardless of speaker, accent, and pitch, all while doing so in the presence noise or over degraded acoustic channels. Furthermore, the system is dynamic and can adapt itself as the acoustic environment or listening objectives change: the same mechanisms that allow us to focus on a particular melody line in an orchestra also help us better follow the speech of a friend at a noisy cocktail party.

Because the auditory system is inherently noise robust and can adapt itself as conditions change, it serves as a useful reference for developing signal processing strategies to extract information from sound. This is especially important because the proliferation of automated sound processing applications on mobile devices, ranging from digital personal assistants that recognize speech to those that identify songs playing in the background, has greatly expanded the need for signal processing strategies that operate in the presence of noise and across a variety of acoustic environments.

This thesis concerns computational models of the signal processing and neural coding strate-

gies at work in the auditory system, with a focus on processing at the level of primary auditory cortex. In particular, we consider models that seek to explain (1) how the system forms a representation that reflects the statistics and structure of natural sounds, (2) how these representations can be used to extract information from sound that is invariant to noise, and (3) how the representation adapts itself as acoustic conditions and listening objectives change. In the course of this thesis, we show not only how our models contribute to a better understanding of the computational mechanisms operating in the central auditory system, but how they can inform signal processing strategies for automated sound processing tasks.

1.1 Signal Processing in the Ascending Auditory Pathway

We begin with a basic overview of the relevant neurophysiological components and corresponding signal processing models that form the foundation of this thesis. A comprehensive review of the full neurophysiological details can be found in [1], and a basic overview of the signal processing pipeline describing the ascending auditory pathway is shown in Fig. 1.1.

1.1.1 Peripheral and Midbrain Processing Stages

Auditory peripheral processing refers to those stages responsible for transducing sound observed at the outer ear to a series of neural firing patterns in the auditory nerve beyond the cochlea. The basic process is as follows. Sound travels down the ear canal and vibrates the tympanic membrane, which in turn displaces small bones in the middle ear (the ossicles) that cause vibrations at the oval window of the cochlea, a spiral shaped, bony structure in the inner ear. These vibrations cause displacement of fluid in two tubes in the cochlea, with a concomitant displacement of the basilar membrane that separates the two tubes and runs along the length of the cochlea. Variations in the thickness and width of the basilar membrane result in high frequency sounds causing maximum displacement at its base, whereas low frequency sounds cause maximal displacement at the apex.

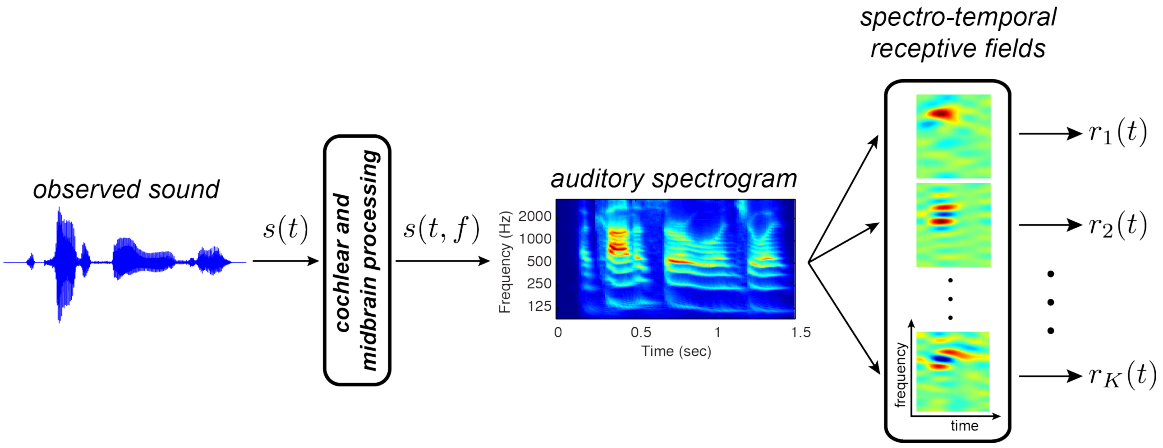

Figure 1.1: Signal processing in the ascending auditory pathway.

This results in an approximately logarithmic frequency-to-place ("tonotopic") mapping of sounds at different points along the basilar membrane. Along the length of the basilar membrane are bundles of hair cells (inner hair cells), which, when sufficiently displaced, open gated ion channels that allow in cochlear fluid that depolarizes the cells, ultimately resulting in the generation of action potentials in the auditory nerve.

Midbrain processing refers to further analysis beyond the auditory nerve by the cochlear nucleus (CN) through approximately the inferior colliculus (IC). These stages preserve the tonotopy instantiated in the cochlea and play a number of important roles including sharpening of neural responses, detecting harmonicity, and integrating spatial cues.

From a signal processing perspective, the time-varying frequency-to-place mapping performed by the cochlea resembles a short-time Fourier analysis of the observed sound, and has inspired a biomimetic representation of sound referred to as an *auditory spectrogram* [2]. For the purposes of this thesis, an auditory spectrogram is generated by first convolving the observed waveform $s(t)$ with a bank of 128 constant-Q bandpass Gammatone-like filters $g_f(t)$. The filters are spaced uniformly along the tonotopic axis and span 5.3 octaves, starting at 90 Hz. This is followed by a first-order derivative and half-wave rectification to model lateral inhibition in the CN and effectively sharpens the filter responses. Lastly, short-term integration in the midbrain is modeled by convolving these

responses with a lowpass filter $\mu(t;\tau) := \mathrm{e}^{-t/\tau}u(t)$ where $u(t)$ is the unit-step function. Typical values for τ in this thesis are on the order of 10 ms. Mathematically, we can compactly summarize these various processing steps as

$$s(t,f) = \max\left[\frac{\partial}{\partial f}\left(s(t) *_t g_f(t)\right), 0\right] *_t \mu(t;\tau)$$

where $*_t$ denotes convolution in time, and an implementation of these steps is available in [3]. An example of an auditory spectrogram for a speech waveform is shown in Fig. 1.1. It exhibits all the hallmarks of a typical spectrogram for speech: energy at the fundamental frequency corresponding to the speaker's pitch, harmonic stacks that indicate voicing, time-varying formants that carry phonemic information, and bands of high-frequency energy corresponding to unvoiced speech sounds.

1.1.2 Central Processing Stages

Beyond the IC, the representation of incoming acoustic stimuli becomes considerably more complex and abstracted beyond the raw sensory input [4]. In fact, it is believed that by the level of the medial geniculate body (MGB) and primary auditory cortex (A1), the brain begins the formation of auditory objects [5], perceptual constructs that correspond to the sounds assigned to a particular source [6, 7]. Furthermore, because neurons at the level of A1 tend to largely be strongly driven by sounds with rich spectro-temporal dynamics [8], it is therefore important to have a useful characterization of a neuron's behavior in response to these kinds of sounds.

One important functional descriptor of a neuron's response is its *spectro-temporal receptive field* (STRF). The STRF has a rich history in auditory neuroscience [9–11], and has proven useful not only for describing basic processing aspects of auditory neurons (e.g., [12–15]), but also for shedding light on the nature of top-down, attention-driven plasticity (e.g., [16–18]). The STRF is a linear characterization of the mapping between a stimulus and an observed neural firing rate, and can be thought of as a matched filter that acts on an input auditory spectrogram. Examples of

STRFs commonly found in mammalian auditory cortex are shown on the right side of Fig. 1.1 [16]. One can observe that the filters are sensitive to various patterns in the input, including localized and fairly narrowband bands of energy, changes along the spectral axis, and joint changes along the spectral and temporal axes.

We can express the firing rate $r(t)$ of a neuron by denoting the STRF as a filter $h(t, f)$ that acts on an observed spectrogram stimulus $s(t, f)$ as

$$r(t) = \int h(t, f) *_t s(t, f) df + r_0 \tag{1.1}$$

where $*_t$ denotes convolution in time and integration is along the frequency axis, and r_0 is the neuron's baseline response. While this does not predict the exact spiking pattern of a neuron, $r(t)$ can be interpreted as the rate parameter that drives an inhomeogenous Poisson spiking process [19,20]. Firing rate is typically estimated using the peri-stimulus time histogram (PSTH), which simply counts the number of spikes that occur over a short interval into time-varying bins. Popular methods to estimate STRFs are based on spike-triggered averaging and spectro-temporal reverse correlation [10, 11, 13, 15].

As a mathematical characterization of neural processing, the STRF is attractive due to its intuitive matched filtering operation, which is familiar to those who have studied linear time-invariant (LTI) systems. More importantly, however, the STRF has a fundamental connection with the Volterra series characterization of a nonlinear system, which is a common representation for biological systems [13, 21]. Briefly, the system response r of a stimulus s can written as a sum of (possibly infinite) functionals

$$r = K_0[s] + K_1[s] + \cdots + K_n[s]$$

The Volterra series expansion prescribes the form of the functionals as

$$K_i[s(t)] = \int \cdots \int v_i(\tau_1, \cdots, \tau_i) s(t - \tau_1) \cdots s(t - \tau_i) d\tau_1 \cdots d\tau_i$$

where $v_i(\tau_1, \cdots, \tau_i)$ is referred to as the i'th Volterra kernel. For example, the standard LTI response of a filter can be written using the first-order Volterra kernel as

$$K_1[s(t)] = \int v_1(\tau) s(t-\tau) d_\tau$$

which is as the familiar convolution operation for characterizing the response of an LTI system. It follows that with an appropriate change of variables, one can relate the STRF $h(t, f)$ to the second-order Volterra kernel; for more details, the interested reader is referred to [22].

The STRF is not without its limitations, however. As a linear model it is sometimes a poor predictor of observed neural firing patterns [23], and the diversity of responses that are elicited from different stimulus ensembles hinders its ability to generalize to unseen stimulus conditions [8]. Nevertheless, because of its mathematical foundations, tractability, success in the describing a wide range of auditory neurophysiological behavior, the STRF plays a fundamental role in this thesis. Specifically, it will serve as a crucial tool that allows us to explore the relationship between high-level coding strategies at work in the auditory system and the concomitant sensory transformations that facilitate these computational goals. In the next section, we describe our general approach to modeling coding strategies in the auditory system, and elaborate on some of the specific sensory coding schemes to be considered in following chapters.

1.2 Computational Objectives in Auditory Processing

A central idea in computational and systems neuroscience is to consider the computational objectives that underlie observed neural responses, and it is widely believed that sensory representations are optimized to process the stimuli to which they are exposed in natural environments [24,25]. A general approach for exploring the effects of particular coding strategies in sensory systems is based on *defining* and *optimizing* a statistical objective criterion that quantifies the principle governing the transformation between stimulus and internal representation. Upon convergence, one then compares

the emergent representation to known properties of the sensory system being studied. This is the general approach we adopt in this thesis, and in the next two sections we look to the neurophysiology for inspiration about coding strategies at work in the auditory system.

1.2.1 Coding Strategies and Emergent Representations in the Central Auditory System

A popular hypothesis explored in recent years assumes that neural populations optimize a sparse code. This means that at any given time, only a small subset of a neural population fires to encode a given stimulus [26]. Such a representation is attractive for reasons of coding efficiency (see, e.g., [24]) and conservation of physiological resources [27]. The sparse coding hypothesis has enjoyed particular success in studies of vision (e.g., [28, 29]), and has also been supported more recently by both neurophysiological [30, 31] and computational studies [32–34] of the auditory system.

In the auditory system, it also been observed that some neurons, when driven by their preferred stimuli, exhibit *sustained* firing rates. Measuring from auditory thalamus and primary auditory cortex, Wang *et al.* observed that sustained responses were not simply phase-locked to the fast dynamics of the stimulus, suggesting that this rate-based code represented a meaningful, non-isomorphic transformation of the stimulus [35,36]. Indeed, such a code is particularly important for audition since it directly addresses the issue of how to indicate the continued presence of a sound in a complex acoustic environment. It has also been suggested that sustained responses play a role in auditory scene analysis, forming part of the neural basis for the perceptual restoration of foreground sounds against a cluttered background [37]. Moreover, Wang has argued that a rate-based representation is critical for matching fast temporal modulations present in natural sounds to slower rates found in higher cortical areas [38]. Slower dynamics in acoustic signals are believed to be the main carrier of information in speech and music [39]; are commensurate with temporal dynamics of stream formation and auditory grouping [40]; and may play an important role in multi-modal sensory integration [38]. Related computational studies in vision have suggested how this principle

may underlie the shapes of simple and complex cell receptive fields in primary visual cortex [41, 42]. Importantly, a sustained firing rate, i.e., one that is persistent and therefore slowly changing over time, is related to slow feature analysis, a well known method for extracting invariances from sensory signals [43]. To the best of our knowledge, however, there are no computational studies that explicitly consider the implications of a sustained firing-based code in central auditory areas.

At first glance, the two coding schemes are seemingly at odds: on the one hand a sparse code seeks to minimize the activity of a neural population whereas a sustained firing-based code requires that neural responses persist over time but still form an efficient representation of the stimulus. However, it appears that central auditory responses can strike a balance between the two strategies, with a large, transient population response at the onset of a sound, and a sparse subset of preferentially driven neurons exhibiting a strong, sustained response throughout the sound's duration [38, 44]. This picture suggests a mechanism for detecting and tracking target sounds in noisy acoustic environments and for generating a persistent signal that facilitates a stable perceptual representation. From a computational perspective, a better understanding of these mechanisms can inform models of auditory scene analysis as well as signal processing schemes for hearing prosthetics and automated sound processing systems like those considered in this thesis.

1.2.2 Top-Down Attention and Plasticity in the Auditory System

It is next important to consider that carefully designed objective functions, while potentially explaining emergent behavior by neural systems, must work in concert with known adaptation mechanisms that allow a system to optimally perform as environmental statistics and behavioral objectives change. This ability, referred to as neural *plasticity*, is a ubiquitous property of sensory cortex whereby neural tuning characteristics can be dynamically shaped based on expectations, environmental context, and behavioral demands. Rapid plasticity has been documented across many sensory modalities including vision [45], somatosensation [46], olfaction [47], and audition [48]. A particularly important driver of neural plasticity is top-down attention, which acts to adapt cognitive

resources to selectively focus on behaviorally relevant sensory input. Such mechanism helps sensory systems dynamically parse the flood of incoming stimuli as environmental context and behavioral demands change over time. For example, attention helps guide the visual search for a friend in a crowd, or it can help a listener follow a specific voice in a cocktail party.

Broadly speaking, attention is a multifaceted and distributed process. Its effects are manifested neurophysiologically at various levels in the cortical hierarchy [16, 49–52], cognitively at many levels of abstraction of the raw sensory input [53, 54], and are dependent on factors such as stimulus statistics [55], task difficulty [18], and the physical constraints of the underlying neural circuitry [56]. Nevertheless, a common computational goal can be identified from studies of top-down attention across sensory modalities: that neural tuning characteristics adapt to improve discrimination and separation between the representation of the foreground (i.e., the attended stimuli) and that of the background (i.e., task-irrelevant distractors).

Studies of attention-driven plasticity have a rich history in the visual domain [57–59]. Neurophysiological studies have described a number of neural parameters that are modulated by attention to facilitate foreground/background separation, including response gain [60], feature tuning bandwidth [61], preferred spatial location [62], and contrast response functions [63]. Furthermore, these observations can be explained by a plethora of computational models [64, 65]. Early connectionist models describe how attention acts to adapt synaptic weights in a distributed neural network to attend to, and emphasize the representation of, desired spatial locations or features [66, 67]. More recent efforts have proposed frameworks that unify a variety of attention-driven effects observed in neurophysiological studies, quantifying how attention acts to bias the gains and/or feature tuning functions of neurons to emphasize target-specific features while suppressing the responses to task-irrelevant features [68–70]. Overall, these models have been important for establishing a theoretical foundation on which to base questions of the optimal computational strategies for, and the neural substrates of, top-down attention, as well as the meaning, interpretation, and scope of top-down signals [71].

In the auditory system, recent neurophysiological studies have begun to shed light on the nature of the computational principles underlying attention-driven plasticity [72–74]. Along the central auditory pathway, top-down attentional mechanisms have been shown to dynamically reshape neural tuning characteristics in order to maximize performance of behavioral tasks. These task-driven changes have been summarized by the *contrast filtering hypothesis*, which states that attention acts to enhance representation of attended sounds in the acoustic foreground relative to those in the acoustic background [75]. Particularly striking examples of contrast filtering effects have been observed in primary auditory cortex (A1) via measurements of STRFs. Specifically, it has been shown that STRFs adapt to directly enhance individual acoustic features of the foreground while suppressing those of the background, and, importantly, that the direction of plasticity reflects the structure of the task and behavioral meaning assigned to foreground and background stimuli [16, 17, 75–77]. Moreover, despite being subject to dramatic changes in their shape, STRFs exhibit remarkable stability in their tuning characteristics by resisting change over time and/or returning to their nominal shapes post behavior [78]. Furthermore, contrast filtering effects have been observed beyond A1 in secondary auditory belt areas up through executive control areas in prefrontal cortex [52, 79]. Thus, the computational principles underlying task-driven plasticity can be understood through the lens of a contrast filter that allows the auditory system to dynamically reallocate neural resources in a discriminative fashion to improve performance in specific tasks while maintaining a notion of representational stability over time.

Recent computational modeling efforts have predicted plasticity patterns that are broadly consistent with the contrast filtering hypothesis in A1 [76, 80]. Broadly speaking, these studies propose discriminative cost functions that maximize a notion of distance between neural responses to foreground and background stimuli to determine optimal receptive field parameters subject to biologically plausible constraints. Importantly, these models predict localized differential plasticity effects that reflect the acoustic features of task-relevant stimuli. They are primarily driven by the physical characteristics of the sensory input and represent—by design—models of *feature-based*

attention. Although quite informative about computational strategies underlying A1 adaptation patterns, these approaches are limited in two important ways. First, they do not capture the influence of task structure on the direction of plasticity effects. In particular, recent data from mammalian primary auditory cortex suggest that during a tone vs. noise discrimination task, aversive tasks (where the target tone is associated with a negative reward) tended to *enhance* representation of the tone whereas appetitive tasks (where the target tone is associated with a positive reward) tended to *suppress* representation of the tone [76]. Because the models define quadratic cost functions whose optima will not change if the roles of foreground and background are reversed, they are therefore agnostic to task structure, with no way to guarantee that plasticity predicted by the models will change direction if the behavioral meanings assigned to foreground and background stimuli are exchanged. Second, because the computational models adapt receptive field parameters based directly on the raw spectro-temporal stimulus—and hence the raw features that characterize the acoustic classes—they lack a mechanism to adapt based on abstractions of the stimulus (e.g., spectro-temporal modulation profile, phase profile, etc.), which one would expect from an *object-based* model of attention that defines the target class along certain characteristics but unconstrains others to allow for variability within the target class [7, 54, 81].

1.3 Thesis Overview

Given this broad overview of our approach to modeling neural systems, we proceed with a more detailed description of the content and contributions of this thesis. We begin in Chapter 2 by considering the implications of a sustained firing criterion for governing the mapping between a stimulus and its corresponding neural response. We first demonstrate the emergence of an ensemble of richly structured model STRFs that reflect the diversity and structure of natural sounds. We compare the emergent STRFs to known neurophysiological results, and we consider how the sustained firing criterion relates to other popular sensory coding strategies based on sparsity and

slow feature analysis, both of which are thought to operate in other sensory modalities. We finally illustrate how the emergent STRFs imply a strategy for noise-robust automatic speech recognition.

In Chapter 3, we propose a computational framework for modeling attention-driven adaptation in ensembles of physiological STRFs. The framework defines a discriminative cost function whose optimal solution explains how STRFs adapt to enhance and suppress features of the acoustic foreground and background, respectively. Importantly, the adaptation patterns predicted by the framework have a close correspondence with known neurophysiological data. Furthermore, a generalization of the framework acts on the spectro-temporal dynamics of task-relevant stimuli, and we make predictions for plausible tasks that have yet to be experimentally measured. We argue that our generalization represents a form of object-based attention, which helps shed light on the current debate between feature- and object-based attentional mechanisms.

In Chapter 4, we consider how our proposed generalized object-based model of attention can be applied to the problem of noise-robust information extraction from speech, using speech activity detection (SAD) as a representative task. SAD is a fundamental first step for many speech processing tasks, and its performance is often severely impaired by the presence of noise. We show how use of an attention-modulated STRF ensemble helps increase representational separation of speech and non-speech sounds, resulting in improved speech detection in additive noise environments. To better understand these results, we show, via stimulus reconstruction experiments, how an attention-modulated ensemble of STRFs forms a high-fidelity and noise-robust representation of an attended target in both clean and noisy conditions.

Lastly, we conclude the thesis in Chapter 5, reviewing the main results from the previous chapters, and identify avenues of future research.

Chapter 2

A Representation for Natural Sounds Based on Sustained Firing Rates

2.1 Introduction

The processing characteristics of neurons in the central auditory system are directly shaped by and reflect the statistics of natural acoustic environments, but the principles that govern the relationship between natural sound ensembles and observed responses in neurophysiological studies remain unclear. As we described in Chap. 1.2.1, accumulating evidence suggests the presence of a code based on sustained neural firing rates, where central auditory neurons exhibit strong, persistent responses to their preferred stimuli. Such a strategy can indicate the presence of ongoing sounds, is involved in parsing complex auditory scenes, and may play a role in matching neural dynamics to varying time scales in acoustic signals.

In this chapter, we describe a computational framework for exploring the influence of a code

based on sustained firing rates on the shapes auditory STRFs. We demonstrate the emergence of richly structured STRFs that capture the structure of natural sounds over a wide range of timescales, and show how the emergent ensembles resemble those commonly reported in physiological studies. Furthermore, we compare ensembles that optimize a sustained firing code with one that optimizes a sparse code, another widely considered coding strategy, and suggest how the resulting population responses are not mutually exclusive. We next show how the emergent ensembles contour the high-energy spectro-temporal modulations of natural sounds, forming a representation that captures the full range of modulations that characterize natural sound ensembles. These results, in turn, imply a noise-robust signal processing strategy applicable to an automatic speech recognition task. Finally, we comment on the relationship between the sustained firing criterion and slow feature analysis, which connects our results with a broader sensory processing scheme. Overall, our findings have direct implications for our understanding of how sensory systems encode the informative components of natural stimuli and potentially facilitate multi-sensory integration.

2.2 Methods

2.2.1 Stimulus description and preparation

An ensemble of natural sounds comprising segments of speech, animal vocalizations, and ambient outdoor noises was assembled for use as stimuli. Two sets were generated, one for training and one for evaluating the response characteristics of the STRFs. Phonetically balanced sentences read by male and female speakers were used [82]. Examples of animal vocalizations included barking dogs, bleating goats, and chattering monkeys [83]. The ambient sounds included, for example, babbling creeks and blowing wind, and other outdoor noises. The speech utterances were approximately three seconds each and comprised 50% of the stimulus. The animal vocalizations and ambient sounds formed the remaining 50% of the stimulus (25% each), were broken into three-second segments, and were windowed using a raised cosine window to avoid transient effects. Finally, segments from each

class were downsampled to 8 kHz, standardized to be zero-mean and unit variance, and randomly concatenated to yield a waveform approximately three minutes in overall length, i.e., ∼90 seconds of speech, ∼45 seconds of animal vocalizations, and ∼45 seconds of ambient outdoor noises.

For each audio sample, we extracted auditory spectrograms [2, 84], and the specific model details were described earlier in Chap. 1.1.1. We note here, however, that we slightly modified the implementation so that we sampled the log-frequency axis over six octaves, starting at 62.5 Hz, with ten equally spaced channels per octave and a short-term integration interval of 5 ms, i.e., we obtained a 60 channel spectral vector every 5 ms.

2.2.2 Spectro-temporal receptive fields

We used the STRF to quantify the relationship between a spectro-temporal stimulus and its corresponding response in central auditory areas, obtaining a firing rate as

$$r(t) = \int\int h(\tau, f) s(t - \tau, f) d\tau df + r_0, \tag{2.1}$$

where $h(t, f)$ is an LTI filter that defines the STRF, $s(t, f)$ is a spectro-temporal stimulus, and r_0 is the baseline firing rate. Without loss of generality, we assume $r_0 = 0$. Observe that the mapping represents convolution in time and integration across all frequencies, and we can interpret the STRF as a matched filter that acts on the input auditory spectrogram.

For discrete-time signals and filters, and assuming that $h(t, f)$ has a finite impulse response, we can express Eq. 2.1 compactly in vector notation as

$$r(t) = \mathbf{h}^T \mathbf{s}(t), \tag{2.2}$$

where $\mathbf{s}(t)$, $\mathbf{h} \in \mathbb{R}^d$ are column vectors denoting the stimulus and filter, respectively [85]. Furthermore, to express the response $\mathbf{r}(t) = [r_1(t)\, r_2(t) \cdots r_K(t)]^T \in \mathbb{R}^K$ of an *ensemble* of K neurons, we

concatenate the STRFs into a matrix $H := [\mathbf{h}_1\, \mathbf{h}_2\, \cdots\, \mathbf{h}_K] \in \mathbb{R}^{d\times K}$ and write

$$\mathbf{r}(t) = H^T \mathbf{s}(t). \tag{2.3}$$

From the stimulus auditory spectrogram, we extracted 250 ms spectro-temporal segments once every 5 ms. Each segment was stacked columnwise into a vector $\mathbf{s}(t) \in \mathbb{R}^d$ where $d = 3000$ (i.e., 50 vectors/segment $\times$ 60 channels). A total of $\sim$30k spectro-temporal vectors were extracted from the stimulus. We subtracted the local mean from each segment and scaled each vector to be unit norm [41], and note that this pre-processing was also applied to the test stimulus used for evaluating the STRF response characteristics. Finally, each spectro-temporal input patch was processed by the ensemble of STRFs to yield a population response $\mathbf{r}(t)$.

2.2.3 Optimization

The objective functions considered in this chapter (see Results) require that we enforce constraints that prevent redundancy in the model STRFs. We considered two sets of constraints which we term *response* and *shape* constraints. First, for the *response* constraints, we constrained the STRF ensemble have unit variance and be mutually uncorrelated. To begin, we first note such constraints can be written as

$$\big\langle r_j(t) r_k(t) \big\rangle_t = \mathbf{h}_j^T \big\langle \mathbf{s}(t)\mathbf{s}^T(t) \big\rangle_t \mathbf{h}_k = \mathbf{h}_j^T C_{\mathbf{s}} \mathbf{h}_k = \delta_{jk},$$

which can then be compactly expressed as an ensemble constraint

$$H^T C_{\mathbf{s}} H = I, \tag{2.4}$$

where $C_{\mathbf{s}} := \langle \mathbf{s}(t)\mathbf{s}^T(t) \rangle_t \in \mathbb{R}^{d\times d}$ denotes the sample covariance matrix and $I \in \mathbb{R}^{K\times K}$ is the identity matrix. Since $C_{\mathbf{s}}$ is real-symmetric, it is unitarily diagonalizable as $C_{\mathbf{s}} = E\Lambda E^T$, where

$E \in \mathbb{R}^{d \times d}$ is a matrix of (columnwise) eigenvectors with corresponding eigenvalues along the diagonal of $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d) \in \mathbb{R}^{d \times d}$. Substituting this decomposition into Eq. 2.4, we obtained

$$\begin{aligned} H^T C_{\mathbf{s}} H &= H^T (E \Lambda E^T) H \\ &= H^T E \Lambda^{1/2} \Lambda^{1/2} E^T H \\ &= U^T U, \end{aligned}$$

where $U := \Lambda^{1/2} E^T H \in \mathbb{R}^{d \times K}$. By recasting the constraints, we can rewrite the original matrix of STRFs as $H = E \Lambda^{-1/2} U$ and consequently

$$\mathbf{r}(t) = H^T \mathbf{s}(t) = U^T \Lambda^{-1/2} E^T \mathbf{s}(t) = U^T \mathbf{z}(t),$$

where $\mathbf{z}(t) := \Lambda^{-1/2} E^T \mathbf{s}(t)$ corresponds to a *whitening* of the input acoustic data, i.e., $\mathbf{z}(t)$ has a spherical covariance matrix. For computational efficiency, we reduced the dimensionality of the input using a subset of the principal components of the stimulus, i.e.,

$$\mathbf{z}(t) \approx \Lambda_m^{-1/2} E_m^T \mathbf{s}(t),$$

where Λ_m and E_m, $m < d$, are the matrices of eigenvalues and eigenvectors, respectively, that captured 95% of the variance of the input. In this work, we found $m = 468$. Therefore, the core problem we wished to solve is:

$$\arg\max_U J(U) \quad \text{subject to} \quad U^T U = I, \tag{2.5}$$

where $J(\cdot)$ corresponded to either the sustained firing or sparse coding objective function (see Results).

To optimize this nonlinear program, we used the gradient projection method due to Rosen,

the basic idea of which is as follows [86,87]. Let $U_{(n)}$ denote the n'th update to the matrix of (rotated and scaled) STRFs U, let $\alpha > 0$ be a learning rate, and let $m \in \mathbb{N}$ be an integer used to adjust the learning rate. Assume $U_{(n)}$ is a matrix with orthonormal columns that is a feasible solution to the problem in Eq. 2.5. We updated U via gradient ascent as follows:

$$U_{(n+1)} = \mathcal{P}\left(U_{(n)} + \frac{\alpha}{2^m} \cdot \frac{\partial J(U_{(n)})}{\partial U}\right) \tag{2.6}$$

where $\mathcal{P} : \mathbb{R}^{d\times K} \rightarrow \mathbb{R}^{d\times K}$ is a projection of the gradient update so that $U_{(n+1)}$ satisfies the orthonormality constraint required in Eq. 2.5. If the update was such that $J(U_{(n+1)}) < J(U_{(n)})$, we set $m \leftarrow m + 1$ and recomputed the projected gradient update, repeating until $J(U)$ was non-decreasing. Finally, learning ceased when the relative change between $J(U_{(n)})$ and $J(U_{(n+1)})$ fell below a threshold η or a maximum number of iterations were reached; in our experiments, we stopped learning for $\eta < 0.1\%$ or a maximum number of 30 iterations. Upon convergence, the desired STRFs were obtained using $H = E\Lambda^{-1/2}U$. Note that for the case of the sustained firing objective, J_{sus} was formed from the sum of K independent terms, allowing us to directly sort the emergent STRFs according to their contribution to the overall objective function; such a sorting was not possible for the sparsity objective.

Of course, the above procedure requires a suitable projection $\mathcal{P}(\cdot)$, and one was derived as follows [88]. In general, for a matrix $A \in \mathbb{R}^{m\times n}$, we wish to find a matrix $V \in \mathbb{R}^{m\times n}$ with orthonormal columns that minimizes

$$||A - V||_F^2 \quad \text{subject to} \quad V^T V = I.$$

Introducing a symmetric matrix of Lagrange multipliers $L \in \mathbb{R}^{n\times n}$, and recalling that $||A||_F^2 = \mathrm{Tr}(AA^T)$, we sought to find a stationary point of the Lagrangian

$$l(V, L) = \mathrm{Tr}\left[(A - V)(A - V)^T\right] + \mathrm{Tr}\left[L(V^T V - I)\right].$$

Computing the (elementwise) partial derivative of $l(V, L)$ w.r.t. V and setting it to 0 we obtained [89]

$$A = V(I + L).$$

Observing that

$$A^T A = (I + L)V^T V(I + L) = (I + L)^2,$$

we have that

$$(I + L) = (A^T A)^{1/2}.$$

Assuming A had full column rank, then an optimal orthogonal matrix that minimized $||A - V||_F^2$ which can be used for the projection in Eq. 2.6 was found as

$$\mathcal{P}(A) = V = A(A^T A)^{-1/2}. \tag{2.7}$$

Lastly, for the *shape* constraints, we optimized a given objective function subject to the STRFs being orthonormal, i.e., $\mathbf{h}_j \mathbf{h}_k = \delta_{jk}$, we solve

$$\arg\max_H J(H) \quad \text{subject to} \quad H^T H = I.$$

Here we can again use Rosen's projected gradient method in Eq. 2.6 along with the projection defined in Eq. 2.7, but the only difference from before is that it *does not* require pre-whitening of the stimulus.

2.2.4 Characterizing individual STRFs

As we will see, optimizing a particular objective function $J(\cdot)$ yields ensembles of STRFs with many well-structured and interesting shapes. To describe these shapes, we considered measures that described individual spectro-temporal and modulation tuning.

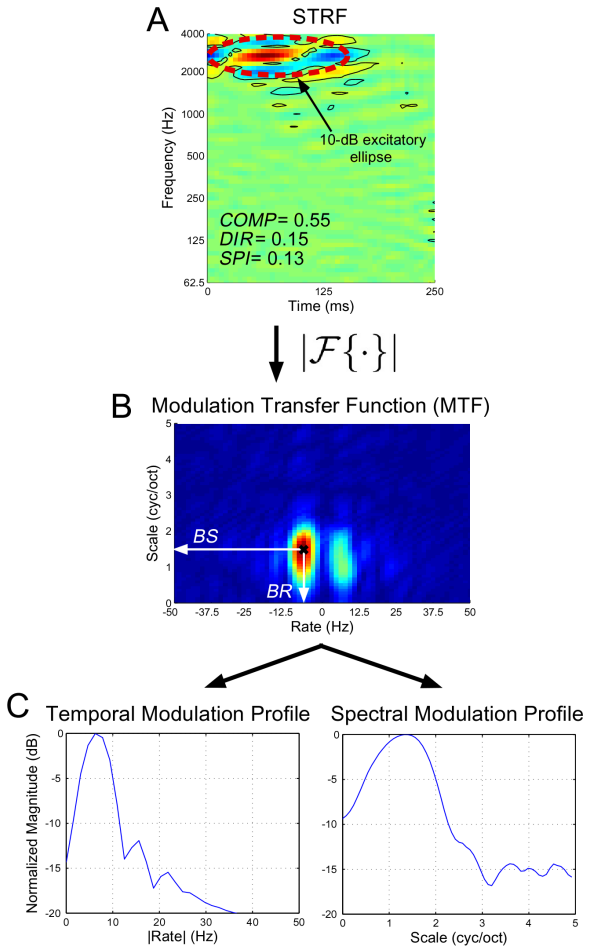

Figure 2.1: Extracting basic spectro-temporal parameters for an individual STRF. Panel (A) shows a typical STRF, with solid contour lines indicating those regions that exceed ± one standard deviation. The dashed red line shows the projected 10-dB ellipse from which we estimated spectral bandwidth. As indicated, the STRF is rather elongated with no strong directional preference, and the pattern is highly separable. Panel (B) shows the MTF computed from the magnitude of the 2D Fourier Transform of the STRF in (A); from here we estimate BR and BS. Panel (C) shows the normalized temporal and spectral modulation profiles obtained from the MTF.

Spectral 10-dB excitatory bandwidth: We measured the spectral spread of an STRF by (1) performing a least-squares fit of a single Gaussian envelope to a thresholded STRF, (2) finding the isoline corresponding to a drop of 10-dB from the maximum of the envelope, and (3) projecting this ellipse onto the spectro-temporal plane. This is referred to as the 10-dB excitatory ellipse and is illustrated in Fig. 2.1A. From this ellipse, we directly the estimated the spectral bandwidth BW_F.

Separability index: We used a measure of separability to quantify how well an STRF $h(t, f)$ could be decomposed into a product of purely temporal and spectral functions, i.e., as $h(t, f) = h_T(t) \cdot h_S(f)$ [12]. Generally speaking, by treating an STRF as a matrix $T \in \mathbb{R}^{m \times n}$, separability can be assessed by considering the singular value decomposition of T:

$$T = U \Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are unitary, $\Sigma \in \mathbb{R}^{m \times n}$ is a matrix such that the *singular values* $\sigma_i \geq 0$ lie along the "diagonal", and $r = \text{rank}(T)$. The separability index SPI was defined as

$$SPI = 1 - \frac{\sigma_1^2}{\sum_{i=1}^{r} \sigma_i^2}.$$

If T is nearly rank-1, we expect σ_1 to dominate and consequently SPI is small, indicating that $T \approx \sigma_1 u_1 v_1^T$, i.e., that the STRF is approximately separable as a product of only two functions. It was often the case that STRFs with a simpler structure, e.g., localized or purely spectral, had small values of SPI. More complex STRFs, particularly those that were noisy, had larger values SPI since they were poorly approximated by a low-rank decomposition.

Modulation Transfer Function: To characterize spectro-temporal modulation tuning in the Fourier domain, we computed the *modulation transfer function* (MTF) of an STRF, illustrated in Fig. 2.1B [90]. The MTF was obtained by computing the magnitude of the 2D Fourier transform of a thresholded STRF; here we set all values of the STRF that did not exceed ± 1 standard deviation to zero. The MTF summarizes the joint sensitivity of an STRF to temporal modulations (*rate*, in Hz) and spectral modulations (*scale*, in cyc/oct).

Best spectral and temporal modulation rates: We selected the peak of the MTF to estimate best rate (BR) and best scale (BS). We expected that BR and BS would summarize an STRF's

preference for fast or slow temporal and spectral modulations.

Average Rate and Scale Profiles: By folding the MTF along the $\omega = 0$ Hz axis, we summarized the temporal and spectral modulation sensitivity of the STRF by summing along each axis, yielding rate and scale profiles; these are illustrated in Fig. 2.1C. These profiles can also be averaged across an ensemble of neurons to yield a population rate or scale profile.

Directionality index: To characterize whether a neuron preferred upward vs. downward stimuli, we computed a directionality index by considering the relative difference in spectro-temporal modulation energy in the first and second quadrants in the Fourier domain. This was quantified as [12]

$$DIR = \frac{E_1 - E_2}{E_1 + E_2}$$

where E_1 and E_2 denote the energy in the first and second quadrant, respectively. By convention, $DIR > 0$ indicates a preference for *downward* moving spectro-temporal patches whereas $DIR < 0$ indicates a preference for *upward* moving spectro-temporal patches.

Compactness: To quantify a notion of compactness for an STRF, we used the *isoperimetric quotient*, which considers the ratio of the area of an ellipsoid to its perimeter, i.e.,

$$COMP = \frac{4\pi \cdot \text{area}}{\text{perimeter}^2}.$$

The area and perimeter were computed from the 10-dB excitatory ellipse which was derived by (1) performing a least-squares fit of a single Gaussian envelope to a thresholded STRF, (2) finding the isoline corresponding to a drop of 10-dB from the maximum of the envelope, and (3) projecting this ellipse onto the spectro-temporal plane. The compactness measure describes the degree to which the coverage of an STRF is spherical ($COMP = 1$) versus elongated ($COMP < 1$), and was used

for characterizing localized vs. non-localized STRFs for the purpose of grouping STRF clusters (described below).

2.2.5 Characterizing STRF populations

Next, we considered measures that characterized a variety of ensemble-based spectro-temporal and modulation properties.

Ensemble modulation transfer function: By averaging the MTF obtained from each STRF, we obtained an ensemble MTF (eMTF) that characterized the average spectro-temporal modulation sensitivity of a given ensemble [90]. This representation was used to relate the average modulation tuning of an ensemble to the modulations present in the stimulus.

Median activation of most persistent neurons: In addition to analyzing the shapes of the emergent STRFs, we explored the ensemble firing rate characteristics of the emergent neurons. Using a held-out set of natural stimuli, we measured the activation of a neuron as the length of time a response was maintained above ± 1 standard deviation (over time) for that particular neuron. We sorted each STRF according to its median activation time, and considered the median responses of the top 10% "most persistent" neurons for a given ensemble (as these subsets appeared to vary most across ΔT). The distributions of these activations were then used to study the extent to which enforcing a sustained response was reflected in a neuron's output.

Average population response histogram: In order to compare distributions of population responses across ensembles, we computed averaged response histograms as follows. Upon convergence of a given ensemble, we filtered a held-out set of natural sound stimuli through the emergent STRFs to obtain a population response. At each time t, we computed a histogram of the population

response, and computed the average histogram across the duration of the stimulus. These averaged histograms could then be used to compare the average population response characteristics across ensembles.

When comparing the receptive field ensembles from the sparse and sustained sets, we only included the responses of highly-structured, non-noisy STRFs as determined by the clustering results outlined next. This step was necessary to keep the comparison between objective functions fair since the sparse ensemble was dominated by noisy STRFs. This inclusion criterion resulted in 115 and 347 neurons for the sparse and sustained ensembles, respectively.

For comparison, we also calculated the response histograms for stimuli filtered through the first 400 principal components of the stimulus (see Fig A.1 in the appendix) as well as through a set of 400 random STRFs. Recall that the magnitudes of the emergent STRFs were constrained so that that their responses had unit variance over time. Accordingly, we normalized the responses of the principal components and random STRFs to also have unit variance to make a fair comparison.

2.2.6 Average stimulus 2D modulation profile

To summarize the spectro-temporal modulations present in the natural sound stimulus, we averaged the magnitude of the 2D Fourier transform of 250 ms patches (non-overlapping) of the auditory spectrogram.

2.2.7 Clustering STRFs

The optimization procedure resulted in a set of richly structured patterns that suggested the presence of a number of latent classes whose membership varied with both choice of objective function and correlation interval ΔT. To quantify these variations, we applied the normalized spectral clustering algorithm of Ng *et al.* [91].

We defined the similarity s_{mn} between a given pair of STRFs $h_m(t, f)$ and $h_n(t, f)$ by computing the normalized 2D cross-correlation matrix for arbitrary shifts in time and frequency and

selecting the maximum of the *absolute* value of this matrix, i.e.,

$$s_{mn} = \max_{i,j} |c_{ij}(m,n)|$$

where

$$c_{ij}(m,n) = \frac{\sum_t \sum_f h_m(i,j) h_n(t+i, f+j)}{||h_m(t,f)||_F \cdot ||h_n(t,f)||_F}.$$

Importantly, the absolute value of the cross correlation was used here since we wished to group STRFs regardless of whether they were excitatory or inhibitory. Next, we pooled all STRFs we sought to cluster and constructed a pairwise similarity matrix $S = [s_{mn}] \in \mathbb{R}^{N \times N}$. Viewing S as a fully connected graph with edge weights specified by s_{mn}, spectral clustering finds a partitioning of the graph into k groups such that edges between groups have low similarity whereas edges within a group have high similarity.

Defining the degree matrix $D = \text{diag}(d_1, d_2, \cdots, d_N)$ where $d_m = \sum_n s_{mn}$ and unnormalized graph Laplacian $L = D - S$, the normalized spectral clustering algorithm is as follows:

1. Compute the normalized Laplacian $L_{sym} = D^{-1/2} L D^{-1/2}$.

2. Compute the first k eigenvectors $\{v_1\, v_2\, \cdots\, v_k\}$ corresponding to the largest k eigenvalues of L_{sym}.

3. Let $V = [v_1\, v_2\, \cdots\, v_k] \in \mathbb{R}^{N \times k}$ and form a matrix W from V by normalizing each row to have unit Euclidean norm.

4. Denoting $w_n \in \mathbb{R}^k$ as the n'th row of W, cluster the set of points $\{w_n\}$ using the k-means algorithm to obtain clusters $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_k$.

We clustered the STRFs initially into 12 groups. While this number was necessarily an arbitrary choice, it was found to sufficiently capture variations in population diversity with ΔT. However, we found that (i) three of the resulting clusters could be reasonably labeled as `noisy`,

whereas (ii) two of the resulting clusters could be reliably labeled as `localized` (see Results); merely reducing the number of initial classes did not merge the clusters, but instead blurred distinctions among the other major categories we sought to study. We interpreted `noisy` patterns as those with no obvious spectro-temporal structure and not indicative of any subset of the stimulus.

Merging of the initial 12 classes was achieved by computing the average SPI of STRFs from the initial class labels and ranking the classes in descending order. Indeed, the three `noisy` classes had the highest average SPI and consequently resulted in a group with average SPI greater than 0.5. Similarly, the localized STRFs were typically highly spherical and sorting the initial clusters by $COMP$ resulted in the two `localized` classes to be ranked highest. Consequently, we grouped these two clusters that had an average $COMP$ of greater than 0.69. This resulted in a final cluster count of nine classes.

2.2.8 Comparison with Neural STRFs

To compare our modeling results with the neurophysiology, we obtained ensembles of neural STRFs estimated using TORC [92] and speech stimuli [16, 93]. There were 2145 TORC and 793 speech STRFs, and each STRF was pre-processed to cover 110 ms in time (sampling rate = 100 Hz) and span 5 octaves in frequency (sampling rate = 5 cyc/oct). For the spectral clustering analysis, we subsampled the TORC set by randomly selecting 793 STRFs and combined them with the speech STRFs, yielding a total of 1586 STRFs in the neural data set. In this way, the neural data analysis was not biased towards one stimulus type or the other.

2.3 Results

We defined a sustained neural response as one where firing rate changes relatively slowly and is consequently highly *correlated* over time. In particular, we were interested in the characteristics of ensembles of model STRFs $H = [\mathbf{h}_1\, \mathbf{h}_2\, \cdots\, \mathbf{h}_K]$ that promoted sustained responses over a specified

time interval $[t - \Delta T, t]$. Denoting the response of the k'th neuron as $r_k(t) = \mathbf{h}_k^T \mathbf{s}(t)$, where $\mathbf{h}_k$ is the STRF and $\mathbf{s}(t)$ is a spectro-temporal stimulus, we quantified this principle using the following objective function:

$$J_{sus}(H) := \sum_{k=1}^{K} \int_{\Delta T} \alpha_\tau \big\langle r_k^2(t) r_k^2(t-\tau) \big\rangle_t d\tau, \tag{2.8}$$

where $\langle \cdot \rangle_t$ denotes time average. Observe that $J_{sus}(H)$ represents the sum of correlations between signal energies of the k'th neuron over a time interval defined by ΔT across an ensemble of K neurons. If a neuron yielded a sustained response, then each of the $r_k(t)$ would vary smoothly over the specified interval and we expect $J_{sus}(H)$ to be large. Moreover, choice of the correlation interval ΔT allowed us to directly explore the effect of different timescales on the ensembles H that *maximized* Eq. 2.8. Finally, the weights α_τ were chosen to be linearly decaying for $\tau = 0$ to ΔT, reflecting the intuition that recent activity of a neuron likely has more influence on the current output than the past. Note that these weights could be adapted to specifically model, for example, positive- or negative-monotonic sustained responses observed in physiological studies [36].

Alternatively, we explored an objective function that promoted sparsity. A natural way to induce sparsity in a population code is by enforcing a population response whose firing rate distribution is highly peaked near zero (representing frequent *weak* responses), but has long tails (representing infrequent *large* responses), i.e., a distribution with high *kurtosis* [94]. We quantified the sparsity of a population code using sample kurtosis:

$$J_{sp}(H) = \left\langle \frac{\mu_4(t)}{[\sigma^2(t)]^2} \right\rangle_t \tag{2.9}$$

where $\mu_4(t) = \frac{1}{K} \sum_k (r_k(t) - \bar{r}(t))^4$ is the fourth central moment at time t, $\sigma^2(t) = \frac{1}{K} \sum_k (r_k(t) - \bar{r}(t))^2$ is the population variance at time t, and $\bar{r}(t)$ is the population mean at time t.

For both $J_{sus}(H)$ and $J_{sp}(H)$, the basic problem was to find an ensemble of STRFs that *maximized* the respective objective function subject to constraints that (1) bounded the amplitude of the filter responses and (2) minimized redundancy among the learned ensemble. This was achieved

by enforcing the responses have unit variance and be mutually uncorrelated, i.e., $\langle r_j(t) r_k(t) \rangle_t = \delta_{jk}$ where δ_{jk} is the Kroenecker delta function (see Methods); we refer to these as *response* constraints. These constraints ensured that the responses had a bounded magnitude and that the STRFs did not all converge to the same solution.

2.3.1 Emergence of richly structured STRFs

We optimized both the sustained objective $J_{sus}(H)$ and sparsity objective $J_{sp}(H)$ using an ensemble of natural stimuli comprising speech, animal vocalizations, and ambient outdoor sounds. Each ensemble of $K = 400$ filters was initialized at random using zero-mean, unit variance Gaussian noise, and each STRF covered from 0–250 ms in time and 62.5–4000 Hz along the tonotopic axis.

For the sustained objective, we considered a wide range of correlation intervals from very brief ($\Delta T = 10$ ms) to very long ($\Delta T = 2000$ ms). Examples of emergent STRFs for $\Delta T = 125$ ms are shown in Fig. 2.2A. For the spectro-temporal patches shown, red and blue colors indicate that the presence of energy in a particular spectro-temporal region yields excitatory and inhibitory responses, respectively. We observe a variety of STRFs that are highly localized, sensitive to narrowband spectral and temporal events, oriented, and some that are seemingly noise-like and not convergent to any particularly interesting shape. Importantly, such observations about these basic STRF classes align with those made in a number of previous physiological studies (see, e.g., [12,90,95]). Moreover, coverage of the STRFs appears to span the full time-frequency space. These results suggest that the sustained firing objective may underlie part of the coding strategy used by central auditory neurons.

Shown in Fig. 2.2B are examples of emergent STRFs obtained by optimizing the sparsity objective. Indeed, this particular objective yields STRFs that are highly localized and sparsely distributed, with sensitivity to bandlimited spectral and temporal events. While both objective criteria yield noisy STRFs, it is clear that the sparse ensemble is much more noisy, with a less extensive coverage of the basic sound classes as observed with the sustained ensemble.

Figure 2.2: Examples of emergent STRFs. Shown are STRFs learned by optimizing (A) the sustained objective function $J_{sus}(H)$ for $\Delta T = 125$ ms and (B) the sparsity objective function $J_{sp}(H)$. The examples shown here were drawn at random from ensembles of 400 neurons. The sustained STRFs are shown in order of decreasing contribution to the overall objective function whereas the sparse STRFs are shown randomly ordered. Each spectro-temporal patch spans 0–250 ms in time and 62.5–4000 Hz in frequency. For these examples the dynamic range of the STRFs was compressed using a $\sinh(\cdot)$ nonlinearity.

2.3.2 Ensemble diversity varies smoothly with ΔT

Since the information-bearing components of natural sounds vary concurrently across multiple timescales, it was expected that the structure of STRFs learned under the sustained objective would vary with the correlation interval ΔT. Indeed, inspection of the sustained ensembles for a range of ΔT suggested the presence of a number of latent classes whose membership varied smoothly from short to long correlation intervals. To quantify variations in population diversity over ecologically relevant timescales, we performed unsupervised clustering of the emergent STRFs and studied how class membership changed with objective function and correlation interval.

We pooled STRFs from the sparse ensemble and from the sustained ensembles for $\Delta T = 10$, 25, 50, 125, 250, 500, 1000, and 2000 ms, yielding a total of 3600 STRFs. We then applied normalized spectral clustering to discover latent classes among the pooled STRFs. In general, spectral clustering algorithms require an affinity matrix that specifies pairwise similarities between the objects being clustered. Viewing this affinity matrix as an undirected graph, spectral clustering finds a partition of the graph into groups whose elements have common similarity with one another. A natural measure of similarity between STRFs can be derived from the two-dimensional cross-correlation between pairs of spectro-temporal patches. Such a measure is similar to that considered by Woolley *et al.* [96] and is desirable since it does not depend on subjective choice of spectro-temporal features to use for clustering. In this work, we defined the measure of similarity between pairs of STRFs as the *absolute value* of the maximum value of the two-dimensional cross-correlation matrix; we used absolute value since we wished to group similar STRFs regardless of whether they were excitatory or inhibitory. Furthermore, as the STRFs tended to be distributed with a variety of phases in the input space, we considered cross-correlations for arbitrary time-frequency shifts (see Methods for details).

Results obtained using normalized spectral clustering of the emergent ensembles into nine classes are shown in Fig. 2.3. In the center panel of the figure, a stacked bar chart illustrates the the percentage of STRFs at a particular ΔT assigned to one of nine classes. Different segment colors correspond to each of the nine classes, and segment width is proportional to the number of STRFs assigned to that class. Surrounding the bar chart are examples from six classes that best illustrate how diversity varies with ΔT, namely `noisy`, `localized`, `spectral`, `complex`, `temporal`, and `directional` classes. These labels are qualitative descriptors of each class and not quantitative assessments of the time-frequency characteristics of each category.

Inspection of the cluster groupings reveal rich structural variations over a wide range of correlation intervals. In particular, the STRFs labeled according to the `noisy` class are found to dominate the sparse ensemble, with a large presence in the sustained ensemble for $\Delta T = 10$ ms.

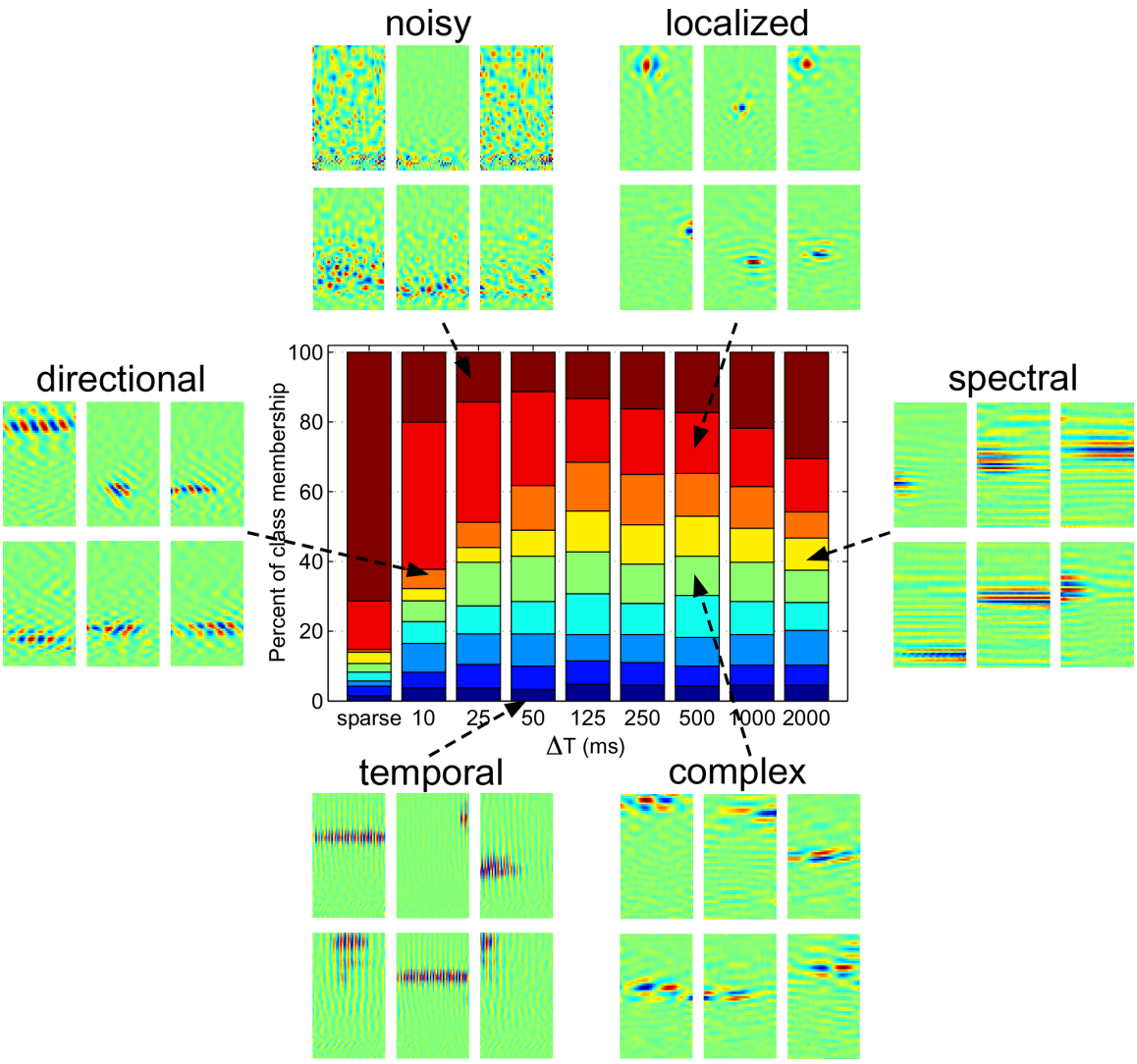

Figure 2.3: Spectral clustering results. Shown are nine clusters obtained by pooling STRFs from the sparse as well as sustained ensembles for ΔT = 10, 25, 50, 125, 250, 500, 1000, and 2500 ms. Shown in the center is a stacked bar chart where segment color corresponds to class label and segment width is proportional to the number of STRFs assigned to a particular class in a given ensemble. The surrounding panels show examples of STRFs drawn from six illustrative classes, namely, `noisy`, `localized`, `spectral`, `complex`, `temporal`, and `directional`.

Membership in this class drops for ΔT between 10 and 125 ms, and begins to increase at 125 ms. We also observe that short correlation intervals (ΔT = 10, 25, and 50 ms) have a large concentration of `localized` STRFs, with membership dropping with increasing ΔT. While the `temporal` class holds relatively steady across the sustained ensembles, we find that membership in the `directional`, `complex`, and `spectral` classes varied smoothly across ΔT. In general, we find that ensemble

diversity is maximized for $\Delta T = 125$ (max. entropy of 3.08 bits), but the overall trends suggest rich ensemble structure between 10 and 250 ms, which is notably in the range of the timescales of natural sounds [97, 98]. This is further supported by the increasing presence of `noisy` STRFs for large correlation intervals ($\Delta T = 1000$ and 2000 ms).

In addition to studying structural variations in the *shapes* of the emergent STRFs, it is also of interest to examine the structure of the STRF *outputs* in response to natural sounds. In particular, we sought to address the extent to which enforcing sustained responses does indeed yield responses that persist over time. We defined the k'th neuron to be significantly "active" when its firing rate $r_k(t)$ exceeded $\pm$ 1 standard deviation over time. While this is not meant to be a precise measure of a neuron's activation (since, for instance, the firing rate is not used to modulate a Poisson spike generation process), such a measure nevertheless quantifies and characterizes a strong versus weak ensemble response to natural stimuli.

Shown in Fig. 2.4A are the distribution of activation times for individual neurons for ensembles of $\Delta T = 10$ and 125 ms in response to a held-out set of natural stimuli. The neurons are shown sorted according to decreasing median activation time, and the interquartile ranges of activation time are indicated by the shaded regions. We observed that the most diversity in median activation times across ensembles occurred in approximately the top 10% of the *most persistent* neurons. To summarize these observations, we considered the distribution of median activation times of the top 10% of neurons with most persistent responses (i.e., the top 40 neurons); these distributions are illustrated as boxplots in Fig.2.4B.

As noted previously with the clustering results, shorter ΔT values favor mostly localized and noisy STRFs and consequently it was expected that activations would be brief. Interestingly, however, we observe that with increasing ΔT, median activations peak between 50 and 500 ms and fall off for large ΔT despite the STRFs being optimized to promote sustained responses over long intervals. This overall trend aligns with the previous clustering results that demonstrate how population diversity is maximized over intervals corresponding to timescales that predominate natural

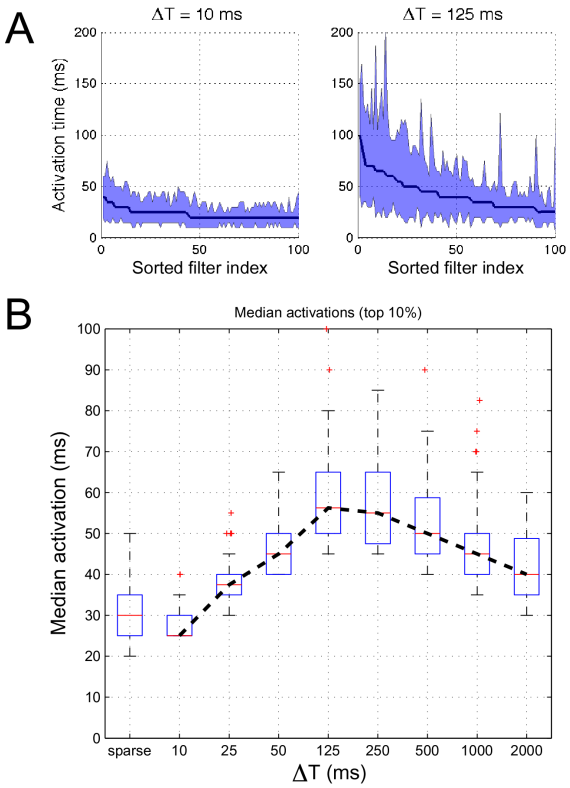

Figure 2.4: Analysis of the temporal activations of emergent ensembles. Panel (A) shows the median activation time of individual neurons (solid lines, sorted in decreasing order) for $\Delta T = 10$ and 125 ms, respectively, for STRFs that optimize the sustained objective function. The shaded region illustrates the corresponding interquartile range. Panel (B) shows the distributions (as boxplots) of median activation times of the top 10% "most persistent" neurons for sparse and sustained ensembles for increasing ΔT.

stimuli. The STRFs corresponding to the top 10% most persistent responses for $\Delta T = 125$ are shown in Fig. A.2 (in the appendices), and we find that they generally have a spectral tuning, but are fairly narrowband and localized.

Additionally, we considered the responses of the top 40 most persistent responses obtained using the sparsity objective function; the distribution of median activations is in the first column of Fig. 2.4B. We find that the sparse ensemble yields responses most similar to those for short ΔT.

2.3.3 Comparison of emergent sustained ensembles to physiology

How do the emergent STRFs learned under the sustained firing objective compare to those observed in physiological studies? Broadly speaking, we find that the emergent STRFs share many of the trends with biological receptive fields typically observed in animal models. We explored this issue by comparing our model ensembles with a set of 1586 STRFs recorded from awake ferret primary auditory cortex using TORC [92] and speech stimuli [16,95] (see Methods for more details). Where applicable, we also compared our results with reported results from anesthetized ferrets by Depireux *et al.* [12] and cats by Miller *et al.* [90] in the literature.

Illustrative examples of the types of STRFs found in the neural data are shown in Fig. 2.5. In particular, we find neural STRFs that are qualitatively similar those found in the `localized`, `complex`, `noisy`, and `directional` clusters shown earlier in Fig. 2.3. Because the temporal and spectral sampling rates used in our model are higher than those used in the physiological data, we did not find good matches with the `temporal` and `spectral` classes.

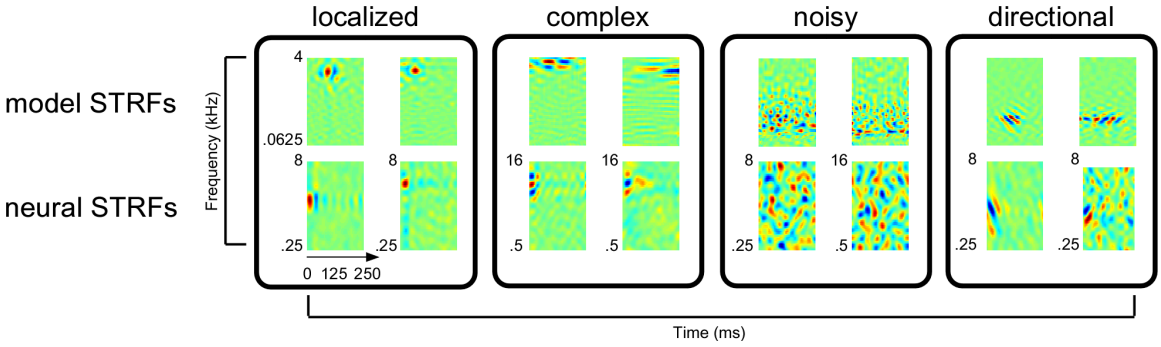

Figure 2.5: Comparison of emergent STRFs learned according to the sustained objective function with examples estimated from ferret auditory cortex.

To visualize the overlap between the spectro-temporal modulation coverage of the neural and model STRFs, we used the ensemble modulation transfer function (eMTF). The eMTF is derived by averaging the magnitude of the 2D Fourier Transform of each neuron in a given ensemble, and jointly characterizes modulations in time (rate, in Hz) and in frequency (scale, in cyc/oct). We first applied normalized spectral clustering to the neural STRFs to obtain nine clusters. Next, we

computed the eMTF for each cluster, extracted isoline contours at the 65% level, and overlaid these curves on the eMTF of the model STRFs for $\Delta T = 125$ ms. These results are shown in Fig. 2.6 and illustrate the overlap between the model and neural data, particularly at the "edges" of the neural STRF modulations. While the overlap is not complete, it is clear that the modulation spectra of each ensemble are not disjoint. Moreover, the model eMTF suggests a general ensemble sensitivity to relatively fast modulations; this point is explored ahead Chap. 2.3.5.

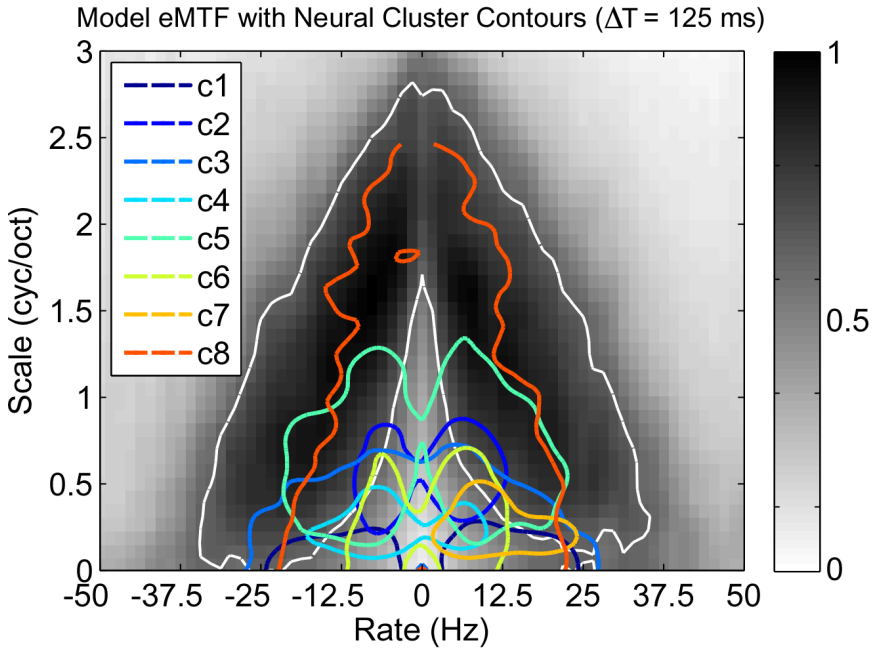

Figure 2.6: Cluster analysis of neural STRFs. Panel (A) shows the overlap between the eMTFs of neural STRF clusters and that of the model STRFs; class 9 comprised mostly noisy STRFs with an exceedingly broad eMTF and its contour is omitted here for clarity. The white contour corresponds to the model eMTF at the 65% level. Panel (B) shows the percent of STRFs assigned to each class by the spectral clustering.

To better characterize the relationship between the neural and model data, we employed a statistical comparison of the distribution of the two datasets. If the models truly generated STRFs similar to those in physiological studies, then one might expect a similarity distribution akin to one derived from the neural ensemble we considered. We computed the symmetric KL-divergence between each of the model and within-physiology similarity distributions (Fig. A.3 in the appendices). We found that the sustained-response (presented here) and sustained-shape (presented

later in this chapter) distributions had KL divergences of 0.80 and 0.85, respectively, whereas the sparse distribution had a KL distance of 1.05. KL typically measures the expected number of bits required to code samples from one distribution using codes from the other. While these numbers are difficult to assess in absolute terms, they give a sense of how the different model optimizations and constraints compare to each other. These numbers reveal that the sustained ensembles are similarly comparable to the physiology, whereas the sparse ensemble has a somewhat worse match. Of course, caution must be taken with any of these numbers because the set of neural STRFs we analyzed represent only a subset of mappings that likely exist in central auditory areas.

Next, we measured a variety of parameters from the neural and model STRFs (for $\Delta T = 125$ ms) that more fully characterized the extent of spectro-temporal coverage and modulation sensitivity of the ensembles, the results of which are summarized in Fig. 2.7.

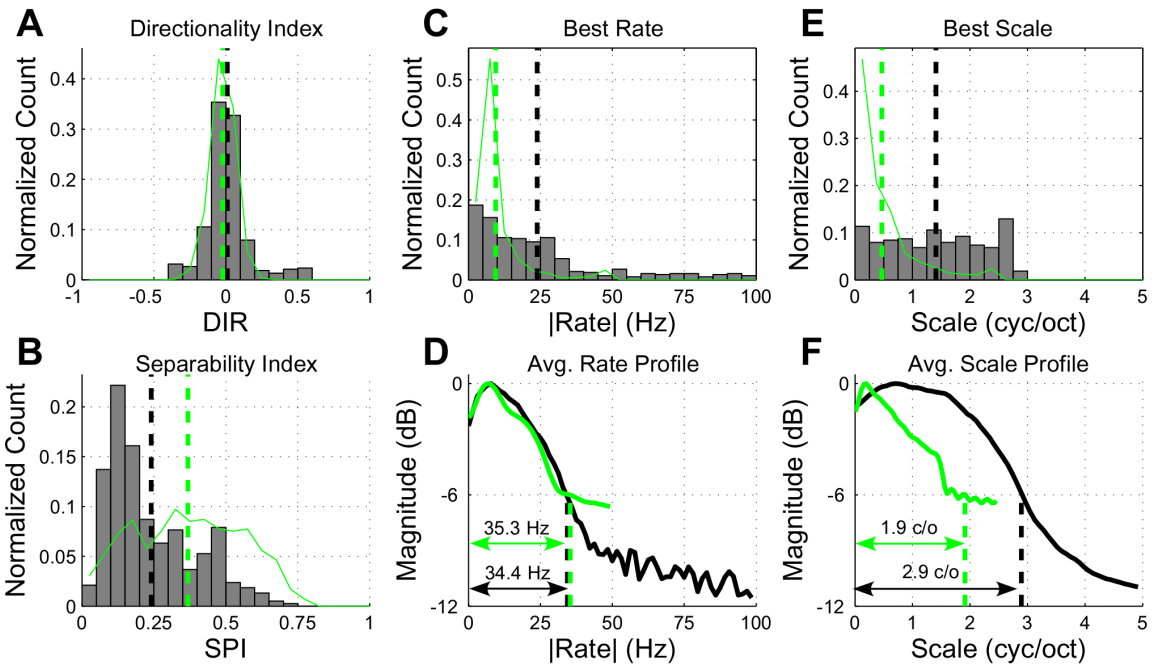

Figure 2.7: Ensemble analysis of STRFs learned under the sustained objective function for $\Delta T = 125$ ms. In panels (A), (B), (C) and (E), the histograms show the distribution of model parameters whereas the black and green dashed vertical lines show population means for the model and neural data, respectively. In panels (D) and (F), the black and green lines correspond to the model and neural STRFs, respectively, with the dashed lines indicating 6-dB upper cutoff frequencies. Refer to text for more details.

Based on the distribution of directionality indices, shown in panel (A), we observe that the model STRFs are largely symmetric, with the majority of neurons having no preference for upward or downward moving input stimuli (mean ≈ 0). As indicated by the tails of this distribution, however, a subset of neurons have a strong directional preference. This agrees with the neural STRFs, and similar observations have been made in MGB and primary auditory cortex of cats by Miller *et al.*, as well as in measurements by Depireux *et al.* from primary auditory cortex of ferrets. Furthermore, panel (B) illustrates that a large number of model STRFs are fairly separable, with a peak in the separability index (SPI) distribution around 0.10 and an average value of 0.26. This trend aligns with values reported in the literature by Depireux *et al.* in measurements from ferret auditory cortex (mean of approx. 0.25). However, it is worth noting that this low level of separability is not uniformly reported across physiological studies of receptive field of mammalian auditory cortex. For instance, the physiological data analyzed in the current study (examples of which are shown in Fig. 2.5) do yield a higher average SPI (mean = 0.37).

The temporal modulation statistics of the model STRFs, as quantified by best rate (BR), also align generally with results reported from mammalian thalamus and cortex. In panel (C) we observe a broad, bandpass distribution of best rates, with an average of 23.9 Hz. Reported physiological results from Miller *et al.* show similarly broad ranges of temporal tuning with preferences around 16 Hz and 30 Hz range for cortex and thalamus, respectively. The neural STRFs we analyzed show a somewhat slower tuning, with an average BR of 9.5 Hz. Furthermore, in panel (D), we computed the normalized average rate profile from the model STRFs. We observe a peak at 7.8 Hz, with an upper 6-dB cutoff of 34.4 Hz. Here we find a close overlap with the rate profile computed from the neural STRFs as well as with average profile results as reported by Miller *et al.* (peak at 12.8 Hz; upper 6-dB cutoff at 37.4 Hz).

The spectral modulation statistics of the model STRFs, as quantified by best scale, are generally faster than those reported from studies of thalamic and cortical nuclei. The distribution of best scales shown in panel (E) is bandpass with a wide range of slow to fast spectral coverage, with

an average tuning of 1.40 cyc/oct. The neural STRFs, in contrast, are tuned to much slower scales (mean = 0.47 cyc/oct). Similarly, results from Miller *et al.* in MGB indicate a generally slower tuning (0.58 cyc/oct), whereas measurements from cortical neurons, while having a similarly wide range of tunings as with the model, indicate a slower average value of 0.46 cyc/oct and an upper cutoff of approx. 2 cyc/oct.

Finally, the ensemble average scale profile, shown in panel (F), is bandpass and exhibits a peak at 0.7 cyc/oct with an upper 6-dB cutoff of 2.9 cyc/oct. The neural STRFs, however, are much slower with peak at 0.2 cyc/oct and an upper cutoff of 1.9 cyc/oct. This is similar to observations from MGB by Miller *et al.*, where they reported that the ensemble average scale profile is generally low-pass, with average scale profile peaks and upper 6-dB cutoffs at 0 cyc/oct and 1.3 cyc/oct, respectively, with similar observations in cortex.

In summary, while we cannot map the emergent STRFs to any exact synapse, they nevertheless reflect the general processing characteristics of various stations along in the central auditory pathway. There is good alignment with the neural STRFs and reported results in mammalian MGB and primary auditory cortex with respect to directional sensitivity and spectro-temporal separability. The temporal modulation statistics of the emergent sustained STRFs appear to be most similar to those measured from thalamus and cortex. Furthermore, the model STRFs are generally faster with regard to spectral modulations than those measured from thalamus and cortex.

2.3.4 Emergence of a sparse population code

To explore the relationship between STRFs optimized to promote sustained responses and those that explicitly maximize population sparsity, we compared the average responses of the sustained ensemble for $\Delta T = 125$ ms with the sparse ensemble. Specifically, we used the converged STRFs to analyze a held-out set of natural stimuli, computed a histogram of the population responses at each time, and computed the average histogram across the entire test input. Since the sparse ensemble was optimized to yield a highly kurtotic firing rate distribution, it was of interest

to examine the shape of the distribution when promoting sustained responses.

Results comparing the average histograms of sustained versus sparse responses is shown in Fig. 2.8, with log-probabilities shown on the vertical axis to emphasize differences between the tails of the distributions. The main observation is that both the sustained and sparse ensembles have distributions that have long tails and are are highly peaked around a firing rate of zero. For reference, we show the average histograms obtained by filtering the stimulus through the first 400 principal components of the stimulus (see Fig. A.1 in the appendices) as well as through a set of 400 random STRFs; a zero-mean, unit variance Gaussian distribution is also shown. Therefore, despite promoting temporally persistent responses, the sustained responses yield a population response that is not altogether different from an ensemble that explicitly maximizes kurtosis. Interestingly, this observation was also made by Berkes and Wiscott in the context of complex cell processing in primary visual cortex (see Sec. 6 of [99]).

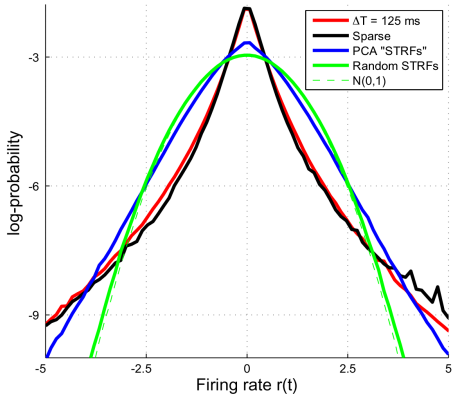

Figure 2.8: Average population response histograms for STRFs learned under the sustained and sparse objectives subject to response constraints.

2.3.5 Emergent STRFs capture spectro-temporal modulation statistics of stimulus

Finally, we sought to explore the consequences of relaxing the constraint that the responses be mutually uncorrelated. Rather than directly constrain the *responses*, we considered constraints

to the *shapes* of the model STRFs. This was achieved by solving

$$\arg\max_{H} J_{sus}(H) \quad \text{subject to} \quad H^T H = I,$$

i.e., we require the STRFs to form an orthonormal basis. So long as the stimuli are bounded, this set of constraints meets our requirements that (1) the output of the STRFs be bounded and (2) we minimize redundancy in the learned ensemble. We consider an ensemble size of $K = 400$ STRFs initialized at random. Examples of shape-constrained STRFs that optimize the sustained objective function for $\Delta T = 125$ ms are shown in Fig. 2.9. Again, we observe STRFs that are bandpass, localized, oriented, and sensitive to a variety of spectral and temporal input. However, there was an apparent difference between the speed of the spectro-temporal modulations and those from STRFs learned subject to the response constraints.

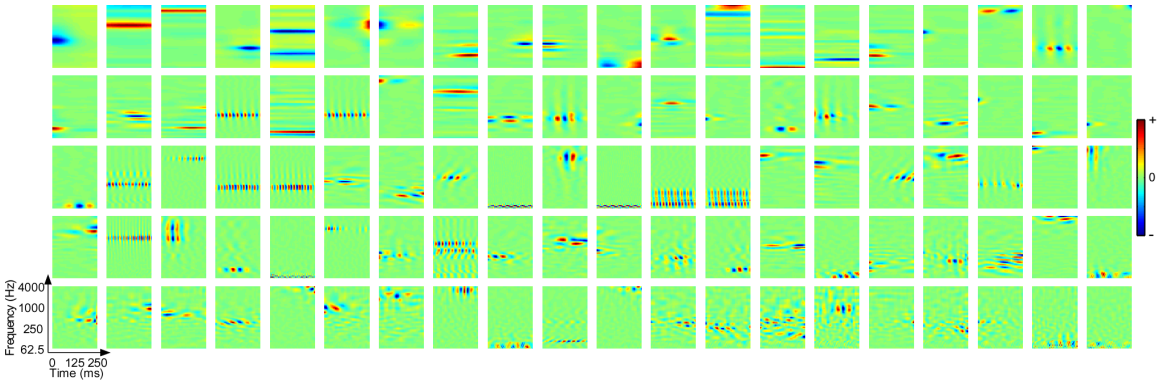

Figure 2.9: Examples of STRFs learned under the sustained objective function ($\Delta T = 125$ ms) subject to orthonormality constraints on the shapes of the filters. The examples shown here were drawn at random from an ensemble of 400 neurons, and the STRFs are shown in order of decreasing contribution to the overall objective function. Each spectro-temporal patch spans 0–250 ms in time and 62.5–4000 Hz in frequency. For these examples the dynamic range of the STRFs was compressed using a $\sinh(\cdot)$ nonlinearity.

It is well known that natural sound ensembles are composed largely of slow spectro-temporal modulations [97, 98, 100]. However, the emergent STRFs learned subject to *response* constraints appear to be tuned to relatively *fast* spectral and temporal modulations, whereas the STRFs learned subject to *shape* constraints appear to have a broader tuning. To further examine how both sets of

constraints jointly capture and are related to the spectro-temporal modulations observed in stimulus, we compared the average 2D modulation profile of the stimulus to the eMTFs derived from both sets of constraints.

An interesting view of how the emergent STRFs capture the spectro-temporal modulations of the stimulus is illustrated in Fig. 2.10 for $\Delta T = 125$ ms. Shown is the average 2D modulation profile of the stimulus overlaid with a single isoline contour (at the 65% level) of the eMTFs learned subject to response (thick red lines) and shape constraints (thick black lines). We also show the constellation of BR versus BS for each ensemble (indicated by '$\times$' and '$\triangle$' for response and shape constraints, respectively). As implied by the contours, the response constraints yield STRFs that follow the spectro-temporal "edge" of the stimulus, while the shape constraints explicitly capture most of the "slowness" of the stimulus. As mentioned previously, the response constraints effectively force the temporal response of the sustained ensemble to be sparse, which consequently results in highly selective STRFs that tend to be tuned to fast modulations. Nevertheless, they implicitly capture the spectro-temporal extent of the stimulus. Moreover, since the shape constraints effectively force the STRFs to form a basis that spans the input space, this results in neurons that explicitly capture the slow modulations of the stimulus. Similar observations were made across the range of ΔT, and for each case it was clear that the spectro-temporal modulations of the stimulus are fully captured by the combination of both sets of constraints.

2.3.6 Application to Noise-Robust Phoneme Recognition

The modulation-domain contouring pattern by the emergent STRFs also suggests a signal processing strategy for automatic speech recognition (ASR). In particular, it is widely believed that "slow" spectro-temporal modulations in speech carry information in a robust manner in degraded acoustic environments [98, 101, 102]. It is therefore expected that signal processing strategies that preserve these slow modulations in noisy environments will yield noise-robust acoustic features. Indeed, application of this principle has been shown by Nemala *et al.* to identify those spectro-

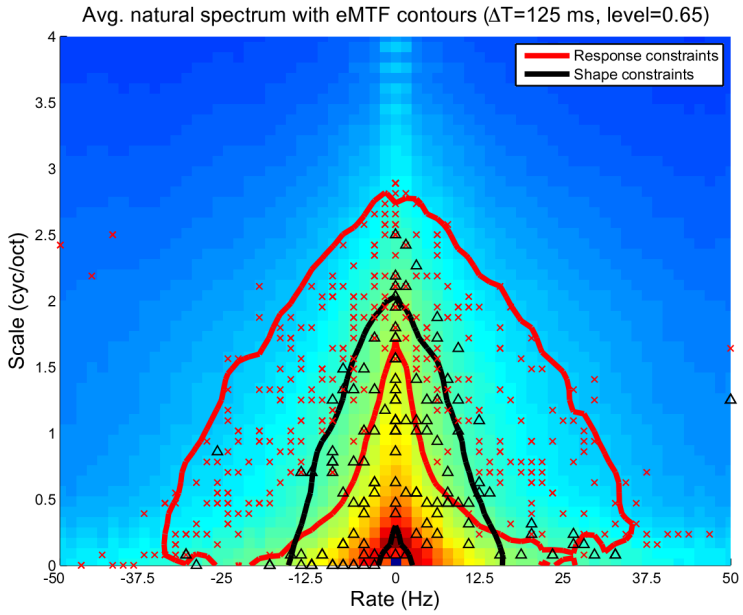

Figure 2.10: Spectro-temporal modulations in the stimulus are fully captured by STRFs that promote sustained responses subject to response and shape constraints. Here, the average MTF of the stimulus is overlaid with contours (at the 65% level) of the ensemble MTFs for both constraints for $\Delta T = 125$ ms. For each ensemble we also show the constellations for best rate vs. best scale (marked by '$\times$' and '$\triangle$' for response and shape constraints, respectively). For the response constraints, we show the contour line and BR/BS constellations for STRFs that contribute to 99% of the objective function.

temporal modulations that yield noise-robust features when corrupted by a variety of additive noise conditions [103]. Here we describe how the sustained firing principle can be used to derive a data-driven 2D spectro-temporal modulation filter for preprocessing auditory spectrograms for noise-robust feature extraction. In a phoneme recognition task, we demonstrate that use of these filtered spectrograms outperform state-of-the-art mean-variance ARMA (MVA) features in both additive noise and reverberant conditions.

Analysis of Emergent STRFs Obtained Using Speech Stimuli

We used approximately three minutes of clean speech from the TIMIT `train` corpus to learn STRFs that optimize the sustained firing criterion subject to response constraints. We used an equal proportion of male and female speakers. Examples of STRFs learned using the above

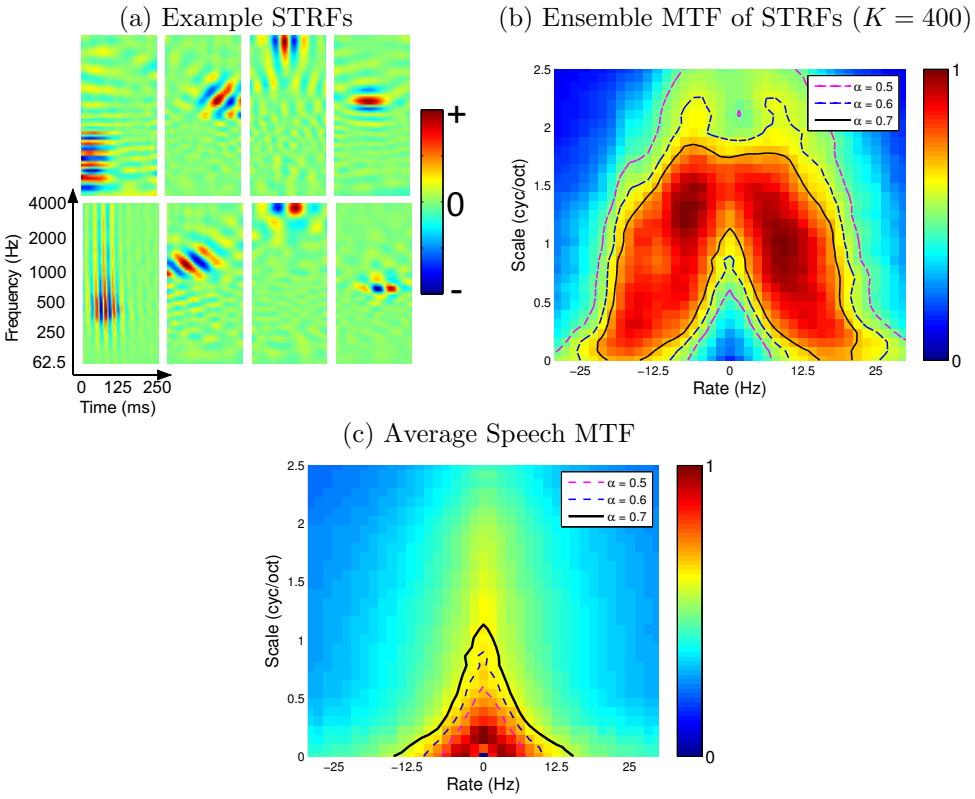

Figure 2.11: Emergent STRFs for use in a noise-robust speech recognition task]. (a) Examples of emergent STRFs learned by optimizing the sustained firing criterion using speech input for $\Delta T = 125$ ms, (b) the corresponding STRF ensemble MTF (eMTF), and (c) an estimate of the average speech MTF. In panels (b) and (c) we superimpose normalized isoline contours derived from the eMTF at various α-levels. For display purposes, the MTF in (c) is compressed by a factor of 1/3.

procedure for $\Delta T = 125$ ms are shown in Fig. 2.11(a). As before, the STRFs exhibit sensitivity to a variety of localized, spectral, temporal, and joint spectro-temporal events in the stimulus.

Shown in Fig. 2.11(b) is the normalized eMTF for an ensemble of $K = 400$ STRFs, again for $\Delta T = 125$ ms. Interestingly, the eMTF shows that the emergent STRF ensemble has little-to-no sensitivity to "slow" modulations (i.e., no energy close to the origin), exhibiting instead a distinct "contouring" effect for rates between approx. $\pm$ 15 Hz and scales between 0 and 2 cyc/oct. It is known, however, that speech has an abundance of modulation energy in these modulation ranges [98], and indeed this is observed when we compute the average MTF of the speech stimulus (Fig. 2.11(c)).

To compare the extent to which the modulation energy of the STRFs contours the modulations of the speech stimulus, we computed normalized isoline contours at the α level (Fig. 2.11(b)),

and considered those portions of the contours closest to the origin (Fig. 2.11(c)). Indeed, when superimposed on the speech MTF, we observe that the contours form a tight boundary around those rates and scales where most of the speech modulation energy is concentrated. This is an especially interesting observation given the recent results of Nemala *et al.* [103], who have demonstrated that auditory spectrograms bandpass filtered to contain only "slow" rates and scales in this region yield noise-robust features for automatic speech recognition.

2D Spectro-Temporal Modulation Filtering

We hypothesize that the observed contouring effect due to the eMTF serves to define the band edges of a 2D bandpass modulation filter in a *data-driven* fashion. The 2D filters we consider are designed directly in the modulation domain using a given contour C at the α level. We set the magnitude response of the filter at rates and scales inside the contour to unity. We then set the roll-off of the filter to be exponential as

$$M_1(\omega, \Omega) = \exp\left\{-\left(\frac{(\omega - \omega_c)^2}{\omega_r} + \frac{(\Omega - \Omega_c)^2}{\Omega_s}\right)\right\} \tag{2.10}$$

where (ω_c, Ω_c) is the point from C that is closest to the point (ω, Ω) being considered. Here, ω_r and Ω_s are the roll-off parameters along the rate and scale axis, respectively. To remove temporal modulations near 0 Hz, we define a wedge function as

$$W(\omega) = \begin{cases} \sin\left(\frac{\pi\omega}{2\omega_W}\right) & |\omega| < \omega_W \\ 1 & \text{otherwise} \end{cases} \tag{2.11}$$

where ω_W is the wedge roll-off along the rate axis [104]. Thus, we obtain the desired 2D filter as $M(\omega, \Omega) = M_1(\omega, \Omega) \cdot W(\omega)$. A given auditory spectrogram is filtered by first transforming to the modulation domain via the 2D Fourier transform, and the magnitude is multiplied with the filter $M(\omega, \Omega)$. Finally, we perform the inverse 2D Fourier Transform, keeping the real part only, to obtain

the filtered auditory spectrogram.

Corpora and Recognizer Setup

Hand-labeled data from the TIMIT corpus was used to train a speaker-independent phoneme recognition system using the Hybrid Multi-layered Perceptron / Hidden Markov Model (MLP/HMM) setup [105]. 3696 utterances were used for training out of which 8% were used as cross validation data. A separate set of 1344 utterances were used for testing. The 61 phoneme labels in the TIMIT corpus were converted to a standard set of 39 labels [106].

A multi-layered perceptron (MLP) was trained discriminatively to estimate the posterior probabilities of the phoneme classes given an input feature vector. The MLP had a hidden layer with 1500 nodes with a sigmoid non-linearity. The output layer consisted of 40 nodes (with a softmax non-linearity) corresponding to the 39 phonemes and an additional garbage class. A second MLP was then trained to include a temporal context of 23 frames (11 frames before and after the current frame) and helped to enhance the posterior probability estimates. The second MLP had the same hidden layer and output layer structure as the first [107].

The HMM system consisted of a three-state feed-forward HMM for each phoneme, with equal probability of transition to itself or the next state. The posterior probabilities were divided by the relative counts of each phoneme and were used as the emission probabilities for the HMM. Finally, the phoneme sequence was decoded using the standard Viterbi algorithm. This decoded sequence of phonemes was compared to the hand-labeled sequence, with recognition rate determined by the number of insertions, deletions, and substitutions.

To assess the noise-robustness of the proposed features, we tested the system under various mismatched conditions. For this we corrupted the test set with additive noise and reverberation. Five types of additive noises from the NOISEX92 corpus [108] were added to the test data at various SNRs from 0–20 dB (at steps of 5 dB) using the FaNT tool [109]. The noises considered were speech babble (Babble), fighter jet cockpit (F16), factory floor (Factory1), military tank (Tank),

and automobile interior (Volvo). For reverberation, we synthesized artificial room responses at five different reverberation time constants (RT_{60}) from 100–500 ms in steps of 100 ms. These responses were generated by convolving Gaussian white noise with an exponentially decaying envelope.

Proposed and Baseline Features

For the proposed features, the auditory spectrogram of each utterance was calculated at a spectral resolution of 24 channels per octave over 5.3 octaves (128 channels in total) at a frame rate of 100 frames/second. We used the contour derived for $\alpha = 0.7$, and 2D filtering (as described in Sec. 2.3.6) was applied with $\omega_r = 1, \Omega_s = 0.12$, and $\omega_W = 1.25$. These constants were empirically determined to maximize performance on the cross validation data set. After applying the 2D filter, we appended first-, second-, and third-order dynamic features, yielding a 512-dimensional input feature vector (i.e., 128×4).

We compared the proposed features with state-of-the-art noise robust features based on MVA processing of MFCC features [110]. These features were obtained by first extracting a standard set of 13-dimensional MFCCs including their first-, second-, and third-order temporal derivatives. Next, cepstral mean subtraction and variance normalization was applied, and the temporal trajectory of each feature dimension was filtered in a RASTA-like manner, further enhancing noise robustness [111]. Finally, a nine-frame context was appended, resulting in a 468-dimensional feature vector (i.e., 13×4×9).

Phoneme Recognition Results

Shown in Table 2.1 are phoneme recognition results for test utterances corrupted by additive noise at a variety of SNRs. It is immediately clear that for clean as well as for all noise types and noise levels the proposed features outperform the baseline MFCC+MVA features, with an overall average absolute gain of 6.4% for the noise cases. This improvement in performance even at 0 dB SNR suggests that the 2D filter is indeed able to capture the high energy regions of speech and

Table 2.1: Phoneme recognition rate (as %) for utterances corrupted by additive noise (higher is better).

Noise Type	SNR (in dB)	Feature	
		MFCC+MVA	2D Filtered
Clean	∞	68.2	69.6
Babble	20	56.6	63.8
	15	49.6	57.7
	10	40.7	47.8
	5	29.8	34.6
	0	19.6	21.8
	Average	39.3	45.1
F16	20	57.1	62.4
	15	50.8	56.5
	10	43.3	47.4
	5	34.6	37.2
	0	27.0	27.2
	Average	42.6	46.1
Factory1	20	55.8	61.6
	15	48.5	55.1
	10	39.5	46.2
	5	30.2	35.6
	0	21.2	25.9
	Average	39.0	44.9
Tank	20	57.8	67.1
	15	54.5	64.7
	10	50.7	60.3
	5	46.4	54.4
	0	41.4	46.5
	Average	50.1	58.6
Volvo	20	63.6	69.6
	15	62.0	69.3
	10	60.2	68.6
	5	58.1	67.2
	0	54.8	64.7
	Average	59.7	67.9

discard the noise regions effectively.

Shown next in Table 2.2 are phoneme recognition results for test utterances corrupted by artificial reverberation. Again, in all cases, we observe that the proposed features outperform the baseline, with an average absolute gain of 3.6%. This further validates the robustness of the filter in capturing the high energy speech regions.

2.4 Discussion

In this chapter, we considered a framework for studying how choice of a sustained firing versus sparse coding objective affects the shapes of model spectro-temporal receptive fields in central

Table 2.2: Phoneme recognition rate (as %) for utterances corrupted by artificial reverberation (higher is better).

Reverb. time (RT_{60})	Feature	
	MFCC+MVA	2D Filtered
100 ms	50.1	53.4
200 ms	37.3	40.6
300 ms	30.5	34.3
400 ms	27.1	30.9
500 ms	24.6	28.3
Average	33.9	37.5

auditory areas. The sparse coding objective considered here, namely that of maximizing population kurtosis, yields STRFs that are mostly noisy. Those that do converge are generally highly localized. In contrast, enforcing the sustained firing objective subject to the same response constraints yields richly structured ensembles of STRFs whose population diversity varies smoothly with the correlation interval ΔT. Of course, the observed structural variations are necessarily biased due to construction of the stimulus. Nevertheless, this diversity, as revealed by the results of the unsupervised clustering, paired with the responses of the most persistent STRFs, supports the notion that sustained neural firings are preferred in the range of timescales predominant in natural sounds. While we do not necessarily attribute the emergent sustained STRFs to any particular synapse in the auditory pathway, we instead note that the observed filters exhibit general similarities to physiological observations made in auditory thalamus and cortex.

We also observed that enforcing the sustained firing objective with response constraints yields an ensemble firing rate distribution that is similar, on average, to one where population sparsity was explicitly enforced. This supports the proposal that the two coding objectives are not necessarily at odds, and that in some sense a sustained firing objective yields "sparsity for free." Of course, the sustained firing and sparse coding objectives could be quantified in many different ways (see, e.g., Hashimoto [112] and Carlson *et al.* [34]), but the present study is a promising step in understanding their relationship in the central auditory system from a computational perspective.

Finally, to explore the consequences of relaxing the constraint that the responses be mutu-

ally uncorrelated, we explored an alternative set of orthonormality constraints on the sustained firing objective. While still minimizing a notion of redundancy, we observed that the emergent ensembles are generally slower, potentially better capturing the slow spectro-temporal modulations known to be present in natural sounds. This experiment further demonstrated the utility of the considered framework for directly addressing questions about coding schemes and various sets of constraints in representing sound in central auditory areas.

2.4.1 Emergence of a discriminative spectro-temporal representation for natural sounds

The combination of shape and response constraints on the sustained objective function yield STRF ensembles that appear to jointly capture the full range of spectro-temporal modulations in the stimulus. However, the distinct differences in MTF coverage illustrate the tradeoff between redundancy and efficiency in sensory representations. In particular, the shape constraints yield STRFs that are somewhat akin to the first few principal components of the stimulus (see Fig A.1). This is not surprising given that the objective function defines a notion of variance of linear projections, the component vectors of which are constrained to form an orthonormal basis. However, since the responses are not strictly enforced to be uncorrelated, orthonormality imposed on the filter shapes does not necessarily reduce redundancy in the resulting neural responses.

In contrast, the response constraints yield STRFs that are highly selective to the input and are thus comparatively "fast" in the modulation domain. This representation can be thought of as more efficient since at any given time only a few neurons have a large response. However, while the shapes of individual STRFs fail to explicitly capture the slow spectro-temporal modulations predominant in natural sounds, it instead appears that the ensemble MTF of the response-constrained STRFs collectively forms a contour around the high-energy modulations of the stimulus that implicitly capture its spectro-temporal extent.

Is this contouring of the average modulation spectrum of natural sounds something per-

formed by the auditory system? The neural STRFs we considered certainly had an eMTF that reflects a tuning to slower modulations near the MTF origin. However, there is some evidence that the auditory system uses an "edge"-sensitive, discriminative modulation profile for analyzing sound. Woolley *et al.* [113], in an avian study, showed that the eMTF of neurons from Field L (the avian A1 analog) has a bandpass temporal modulation profile (at low scales) that facilitates a discriminative tuning of temporal modulations among classes of natural sounds. Nagel and Doupe [114] have also shown examples of avian Field L STRFs that orient themselves near the spectro-temporal "edge" of the stimulus space. Moreover, Rodriguez *et al.* [115], in a study of mammalian IC neurons, showed that neural bandwidths can scale to better capture fast, but less frequent, modulations. In light of these observations, the modulation profiles observed from the sustained STRFs for both response and shape constraints are consistent with the notion that the auditory system makes an explicit effort to capture all modulations present in natural sounds: fast, feature-selective, and consequently *discriminative* modulations, as well as frequently occurring slow modulations.

2.4.2 A neural code for sensory processing

The notion that sustained neural firings form part of the neural representation of sensory systems is not limited exclusively to the auditory modality. In fact, the sustained firing objective considered in this chapter is related to a broad class of sensory coding strategies referred to collectively under the *temporal slowness hypothesis*. This concept proposes that the responses of sensory neurons reflect the time-course of the information-bearing components of the stimulus—which are often much slower with respect to the fast variations observed in the stimulus—and may therefore reflect invariant aspects of the sensory objects in the environment. Examples of early neural network models exploring slowness as a learning principle were considered by Foldiak [116], Mitchison [117], and Becker [118]. More recently, a number of computational studies, particularly in vision, have established slowness as a general sensory coding strategy and have revealed relationships with a number of general machine learning techniques. Here we outline the connections between the sustained

firing criterion considered in this study and previous work.

Our definition of the sustained firing objective, J_{sus}, was adapted from a notion of temporal stability proposed by Hurri and Hyvarinen termed *temporal response strength correlation* (TRSC) [41]. This study considered modeling of simple cells in primary visual cortex, and their objective function was defined as

$$J_{TRSC} = \sum_{k=1}^{K} \langle r_k^2(t) r_k^2(t-\tau) \rangle_t \tag{2.12}$$

for a single fixed τ. By maximizing J_{TRSC} subject to the decorrelation constraints $\left\langle r_j(t) r_k(t) \right\rangle_t = \delta_{jk}$, they showed the emergence of spatial receptive fields similar to those observed in simple cells in primary visual cortex. It is clear that the objective functions J_{sus} and J_{TRSC} are equivalent for a single time step, but the main difference between the two is that we sought to enforce temporal stability over a time *interval* $[t - \Delta T, t]$, rather than between two *distinct* times t and $(t - \tau)$. Interestingly, optimization of the TRSC objective was shown by Hyvarinen to yield a solution to the blind source separation problem [119], suggesting perhaps that in the auditory domain, such a criterion may underlie separation of overlapping acoustic sources.

The sustained firing objective is also related to a well-known model of temporal slowness known as *slow feature analysis* (SFA) [43]. The computational goal of SFA is to find a mapping of an input that extracts the slow, and presumably more invariant, information in the stimulus. Briefly, for an input $\mathbf{x}(t)$, linear SFA finds mappings $y_k(t) = h_k^T \mathbf{x}(t)$ that minimize

$$J_{SFA} := \langle (y_k(t) - y_k(t-1))^2 \rangle_t \tag{2.13}$$

subject to $\langle y_k(t) \rangle_t = 0$, $\langle y_k^2(t) \rangle_t = 1$, and $\langle y_j(t) y_k(t) \rangle_t = 0 \, \forall \, j < k$. Note that the input $\mathbf{x}(t)$ is not necessarily the raw stimulus but could represent a non-linear expansion of the input, akin to applying a kernel function in a support vector machine [120]. Therefore, SFA finds a mapping of the input that varies little over time and whose outputs are bounded and mutually uncorrelated.

In the visual domain, Berkes and Wiskott found that SFA could explain a variety of complex cell phenomena in primary visual cortex such as the emergence of Gabor-like receptive fields, phase invariance, various forms of inhibition, and directional sensitivity [99]. Similar to our study, they also found the emergence of a sparse population code based on SFA. More importantly, however, they established a link between SFA at the level of complex cells and J_{TRSC}, which in turn links to the sustained firing objective J_{sus} explored in our study. Specifically, they showed that when a complex cell output is expressed as a quadratic form $y(t) = \sum_k r_k^2(t)$ [112, 121], the SFA objective could be written as

$$J_{SFA} = \sum_{k=1}^{K} \langle r_k^2(t) r_k^2(t-1) \rangle_t + \sum_{j \neq k} \langle r_j^2(t) r_k^2(t-1) \rangle_t \tag{2.14}$$

which is equivalent to maximizing J_{TRSC} (and thus J_{sus} for a single time-step) plus cross-correlation terms. As noted by Berkes and Wiskott, this relationship suggests that sustained firing rates at the level of simple cells are modulated as part of a hierarchical cortical processing scheme in primary visual cortex. Given the increasing understanding of such hierarchical circuits in the auditory system [4], the possibility that sustained firing rates are varied as part of a higher-order processing strategy in primary auditory areas is an exciting prospect worth further exploration.

Other important relationships exist between SFA and a number of general machine learning principles. Blaschke *et al.* [122] established a relationship between SFA and independent component analysis, a widely used method for blind source separation (see, e.g., [123]). Klampfl and Maass [124] showed that under certain slowness assumptions about the underlying class labels in observed data, SFA finds a discriminative projection of the input similar to Fisher's linear discriminant. Furthermore, SFA has links to methods for nonlinear dimensionality reduction: Creutzig and Sprekeler [125] described the link between SFA and the information bottleneck whereas Sprekeler [126] showed a connection between SFA and Laplacian eigenmaps.

In summary, the temporal slowness hypothesis forms a sound basis for learning a representation from data with rich temporal structure. Slowness as a learning principle has also been

shown to explain the emergence of simple and complex cell properties in primary visual cortex. As described above, the sustained firing principle considered in this chapter has fundamental links to SFA, which in turn is related to a number of general machine learning strategies. To the best of our knowledge, ours is the first thorough study that establishes a link between the temporal slowness hypothesis and an emergent spectro-temporal representation of sound in central auditory areas.

2.4.3 Implications for automated sound processing systems

The ensemble modulation coverage results are particularly interesting since it is widely thought that "slow" spectro-temporal modulations carry much of the message-bearing information for human speech perception. Furthermore, it is known in the speech processing community that preserving slow temporal [127] and joint spectro-temporal modulations [128, 129] is important for noise-robust automatic speech recognition. The observed contouring effect resulting from the sustained firing criterion may thus reflect a mechanism to detect the spectro-temporal "edges" of the message-bearing components of the stimulus, and possibly contribute to a noise-robust representation of sound. Indeed, evidence for this was found when we used a speech-driven MTF contour to design a 2D spectro-temporal filter for preprocessing spectrograms for noise-robust feature extraction. The resulting features outperformed state-of-the-art MVA-processed MFCCs both in clean conditions and in all additive noise and reverberation scenarios considered here.

While we could have derived the filter contours directly from the speech MTF, we consider the question of the information content of spectro-temporal modulations from an alternative but complementary perspective. In particular, the spirit of the work of Nemala *et al.* [103] was to focus resources on subsets of rates and scales that were somehow "linguistically important" and presumably carried the message-bearing components of speech. This was achieved by choosing modulation filter parameters that reflected this intuition in the joint spectro-temporal modulation domain, and is indeed consistent with the RASTA filtering framework of Hermansky and Morgan [111].

It is therefore noteworthy that the sustained firing objective function and associated con-

straints arrive at a similar notion of data-driven filter design. In this work, rather than designing the shape of the modulation filter by hand, we arrived at a noise-robust representation for speech by considering more generally the form of a neural coding strategy used in central auditory areas. Additionally, the emergent neural ensemble, while implicitly capturing the extent of the slow spectro-temporal modulations in the stimulus, primarily exhibits sensitivity to fast modulations relatively far from the origin. Such a distribution may reflect more generally a form of unsupervised learning that discriminates among the various classes of sounds present in speech [113].

Chapter 3

Modeling Attention-Driven Plasticity in Auditory Cortical Receptive Fields

3.1 Introduction

In the previous chapter, we described the *emergence* of STRFs according to a sustained firing criterion, and we studied the relationship between the shapes of the filters, available neurophysiological data, and the statistics of natural sounds. However, it is well known that to navigate complex acoustic environments, listeners *adapt* neural processes to focus on behaviorally relevant sounds in the acoustic foreground while minimizing the impact of distractors in the background, an ability referred to as top-down selective attention. Particularly striking examples of attention-driven plasticity have been reported in primary auditory cortex via dynamic reshaping of STRFs. By enhancing the neural response to features of the foreground while suppressing those to the background, STRFs can act as adaptive contrast matched filters that directly contribute to an improved cognitive

segregation between behaviorally relevant and irrelevant sounds.

In this chapter, we propose a discriminative framework for modeling attention-driven plasticity in STRFs at the level of primary auditory cortex. The model describes a general strategy for cortical plasticity via an optimization that maximizes discriminability between the foreground and distractors while maintaining a degree of stability in the cortical representation. The first instantiation, referred to as the *Feature-Based Model*, describes a form of feature-based attention and yields STRF adaptation patterns consistent with a contrast matched filter previously reported in neurophysiological studies. An extension, referred to as the *Object-Based Model*, captures a form of attention where top-down signals act on an abstracted representation of the sensory input characterized in the modulation domain. The Object-Based Model makes explicit predictions in line with limited neurophysiological data currently available but can be readily evaluated experimentally. Finally, we draw parallels between the proposed framework and anatomical circuits reported to be engaged during active attention. The results strongly suggest an interpretation of attention-driven plasticity as a discriminative adaptation operating at the level of sensory cortex, in line with similar strategies previously described across different sensory modalities.

3.2 Methods

Stimuli and auditory periphery analysis

Stimuli used in the Feature-Based Model included single tones, multi-tone complexes, and spectro-temporally rich broadband noises referred to as a temporally orthogonal ripple combinations (TORCs); these noise stimuli are commonly used to drive neurons in mammalian primary auditory cortex to derive STRFs [15]. We next computed spectrograms as described in Chap. 1.1.1 but to reduce the number of parameters in the optimization described later, the spectral axis was resampled from 128 to 50 tonotopic channels spanning 5.3 octaves. This yielded spectrograms with a spectral sampling rate of 9.4 cycles/octave and temporal sampling rate of 100 Hz.

For the Object-Based model, we considered temporal click-rate and spectro-temporal modulation noise discrimination tasks. For click-rate discrimination, we generated simple click trains by spacing vertical bars in the auditory spectrogram at the prescribed click rate for a given task, and the bars were smoothed in time with a decaying exponential window with a 10 ms time constant. This smoothing helped to spread out temporal modulation energy, rather than having all of the temporal modulation focused solely at the prescribed click rate and its harmonics.

For the spectro-temporal tasks, the stimuli were designed directly in the modulation domain, coupled with random phase, and an inverse Discrete Fourier Transform was performed to obtain the spectrograms in time-frequency; this process is illustrated ahead in Fig. 3.5D and E. We constructed four classes of noise stimuli, referred to as *Narrowband Up* (*NB Up*), *Narrowband Down* (*NB Dn*), *Broadband Up* (*BB Up*), and *Broadband Down* (*BB Dn*). The *BB Up* and *BB Dn* classes shared energy over range of modulations defined by Gaussians centered at ($\pm$16 Hz, 0.5 c/o), and the classes were distinguished by added energy defined by a Gaussian centered at (+16 Hz, 0.25 c/o) and ($-$16 Hz, 0.25 c/o), respectively. The ratio of the Gaussian peaks between target to shared modulations was 2:1. The *NB Up* and *NB Dn* classes were designed similarly, except the shared modulations were centered at ($\pm$10 Hz, 0.5 c/o). The variances of the Gaussians are as specified in Fig. 3.5D.

Auditory cortical receptive fields

We considered an ensemble of 2145 STRFs estimated from recordings from non-behaving ferret primary auditory cortex in response to TORC stimuli [15]. The STRFs spanned 5 octaves in frequency over 15 channels (spectral sampling rate of 3 cycles/octave), with base frequencies of 125, 250, or 500 Hz. Furthermore, the STRFs spanned 250 ms in time over 13 bins (temporal sampling rate of 52 Hz).

We modified the STRFs (1) so that we had finer spectral sampling compared to the original coarse 15 channels of coverage and (2) for convenience so that the frequency range of the STRFs

aligned with the output of the auditory peripheral model. To this end we assumed the base frequency of each STRF to start at 90 Hz, and resampled the spectral axis so that the STRFs spanned 5.3 octaves over $F = 50$ channels. We used cluster analysis (described previously in [130]) to verify that shifting the base frequency of each STRF was not unreasonable since examples from each cluster could be found at each original base frequency (data not shown). We also resampled the temporal axis to span 250 ms over $T = 25$ temporal bins, again to gain finer temporal sampling compared to the original STRFs. Thus, each STRF can be viewed as an image patch $h(t, f) \in \mathbb{R}^{50\times25}$, with a spectral sampling rate of 9.4 cycles/octave and a temporal sampling rate of 100 Hz.

In general, the ensemble formed a richly structured representation of natural sounds, exhibiting sensitivity to localized, spectral, temporal, and joint spectro-temporal acoustic events [14]. We also found that the ensemble contained a large number of "noisy" STRFs, i.e., shapes that appeared unconverged or had no clear preferred spectro-temporal tuning. We used a two-step procedure to remove these noisy STRFs. First, all STRFs were sorted according to the SNR associated with each recording and an initial subset was selected keeping STRFs that had an SNR of at least 2.4 dB. Next, based on the clustering results of Chap 2.2.7, we removed STRFs with $SPI \geq 0.5$. This is because SPI was useful for characterizing clean vs. noisy STRFs, with clean, well-structured STRFs having small SPI and noisy STRFs having large SPI. Overall, application of these two steps yielded an approximately "de-noised" ensemble of 810 STRFs. From this subset, we randomly selected 10 ensembles of size $K = 100$ STRFs and considered these as the initial ensembles $\mathcal{H}_0 = \{h_k^0(t, f)\}, k = 1, 2, \cdots, K$ in our experiments.

Finally, upon ensemble construction, we modeled the notion of a neuron having a finite spectral and temporal integration window by incorporating a spectro-temporal mask in the definition of neural firing rate. For each STRF, a mask was automatically determined by a least-squares fit of a non-oriented Gaussian envelope to a thresholded (at 0.75 standard deviations) and fully rectified STRF.

Optimization and implementation details

As we will describe, because of the nature of the objective functions considered in this chapter, it was not possible to obtain closed-form solutions to the model parameters of interest, necessitating use of numeral optimization techniques. The optimal parameters for the Feature-Based Model were found using the `fmincon` function in the MATLAB optimization toolbox. We used 5 seconds of audio for both the target and reference stimuli. The optimal parameters for the Object-Based Model were determined using `CVX`, a package for specifying and solving convex programs [131, 132]. For this model, we scaled each stimulus token to have unit Euclidean norm, as this seemed to improve optimization convergence. We used 75 tokens, each 250 ms in length, for both the target and reference stimuli. We run each algorithm until the relative change in the objective function is small (threshold of 10^{-6} for the Feature-Based Model and 10^{-4} for the Object-Based Model) or a maximum number of iterations is reached (30 for the Feature-Based Model, 10 for the Object-Based Model).

Feature-Based Model analysis

In line with previous neurophysiological studies (see, e.g., [16]), we quantified the effect of model-induced plasticity on the receptive fields by computing the difference between Euclidean-normalized active and passive STRFs ($\Delta STRF$). This allowed us to directly visualize changes in STRF shape, and $\Delta STRF$ was aligned to the target (or reference) tone frequencies to visualize average population patterns across different tasks. We also derived a measure of relative gain change (ΔA) from the difference STRF at task-related frequency channels. This was computed as the relative change in (normalized) active and passive STRF magnitudes at the location of absolute maximum in $\Delta STRF$ at a particular target or reference channel.

Object-Based Model analysis

For the Object-Based Model, we also considered $\Delta STRF$ defined above to visualize changes between the active and passive STRFs. To visualize model-induced changes in the spectro-temporal modulation profiles, we considered the difference between the modulation transfer functions of the active and passive STRFs (ΔMTF). Average population changes could be visualized in this domain regardless of individual STRF shape and phase [50, 77]. For the click rate discrimination task in particular, all changes in the modulation domain occurred along the rate axis at a scale of 0 cyc/oct due construction of the click train stimuli. For this reason, we considered changes in modulation profile only at this scale in our analysis.

In addition to change in the modulation domain, we sought to characterize STRF changes observed in the time-frequency domain. For spectro-temporal modulation noise discrimination, the model induced clear changes in STRF orientation and directional tuning, so we used the directionality measure DIR to characterize the degree to which a neuron was sensitive to downward vs. upward drifting modulations (see Chap 2.2.4). As a reminder, DIR ranges between $[-1, +1]$, with large positive values indicating sensitivity to downward modulations, and large negative values indicating sensitivity to upward modulations. Finally, to quantify model-induced change in directional tuning, we report the difference in directionality between active and passive settings, defined as $\Delta DIR = DIR_A - DIR_P$. Positive changes in ΔDIR indicate a shift towards sensitivity to downward modulations, and negative changes indicate a shift towards sensitivity to upward modulations.

For click rate discrimination, the model appeared to induce subtle changes in the temporal bandwidth of the STRF main excitatory subfield in the time-frequency domain. We extracted this temporal bandwidth in a simple non-parametric fashion as follows. First, the STRF was interpolated (by zero-padding in the modulation domain) and thresholded at 2 standard deviations to keep significant peaks. Next, the STRF was half-wave rectified and bounding boxes determined for islands of excitatory activity that exceeded threshold. The main excitatory subfield was defined as that which contained the neuron's best frequency / best latency peak, and temporal bandwidth was

defined as the temporal width of the corresponding bounding box.

3.3 Results

3.3.1 Physiological STRF ensemble

We considered ensembles of STRFs obtained from recordings of awake, non-behaving ferret primary auditory cortex; some examples from this ensemble are shown in Fig. 3.1. The STRFs reflect sensitivity to a variety of spectro-temporal events that characterize natural sounds, including localized energy in time-frequency, as well as purely spectral, purely temporal, and joint spectro-temporal modulations. For the experiments described below, we consider ten ensembles of $K = 100$ STRFs randomly sampled (with replacement) from a collection of 810 STRFs.

3.3.2 Overview of the discriminative framework

An overview of the proposed discriminative framework is shown in Fig. 3.2. Broadly speaking, the proposed framework quantifies the physiologically implied balance between discrimination and stability via an objective function of the form

$$J(\mathbf{w}, \mathcal{H}_A) = \text{Discriminability}(\mathbf{w}, \mathcal{H}_A, \mathcal{A}_t, C) + \text{Stability}(\mathcal{H}_0, \mathcal{H}_A, \lambda) \tag{3.1}$$

where $\mathbf{w}$ is a vector of parameters for a discriminative model, $\mathcal{H}_0$ and $\mathcal{H}_A$ are the sets of initial and adapted STRFs, respectively, $\mathcal{A}_t$ is a time-varying *attentional signal* that assigns behaviorally meaningful categorial labels to observed neural responses, and (C, λ) are hyperparameters that control the impact of each term on the overall objective function. In keeping with nomenclature commonly used in auditory physiological studies, we interchange use of foreground with *target* stimuli, as well as interchange use of background with *reference* stimuli. Thus, the overall goal here is to determine settings of $\mathbf{w}$ and $\mathcal{H}_A$ that optimize the proposed cost function.

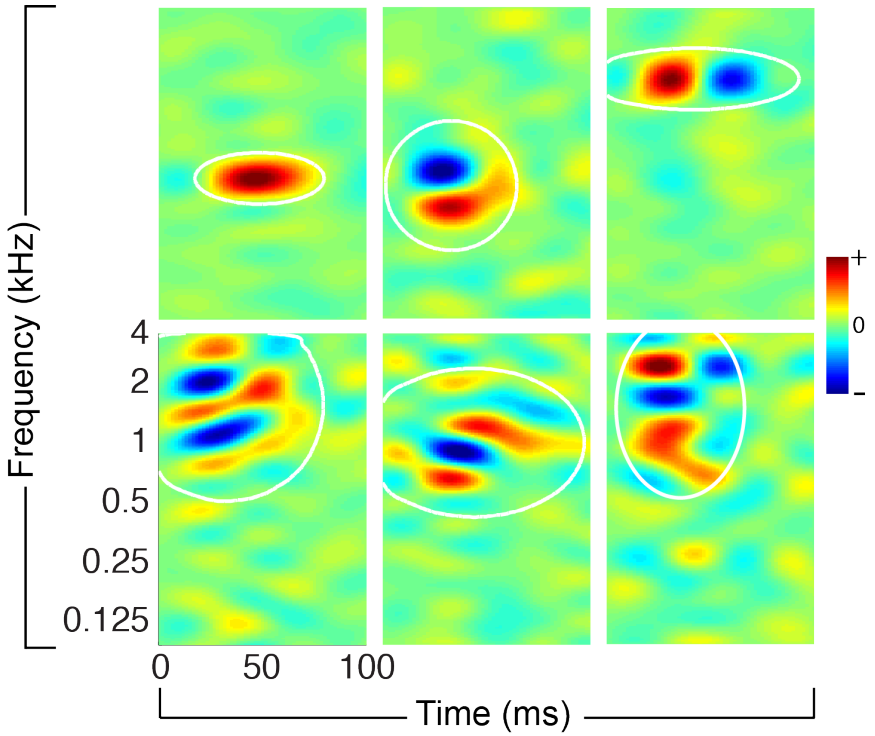

Figure 3.1: Examples of physiological STRFs obtained from mammalian primary auditory cortex. The STRFs reflect sensitivity to a variety of spectro-temporal events in natural sounds, including localized time-frequency energy, spectral and temporal modulations, and more complex joint spectro-temporal modulations. The white ellipses denote isoline contours (at the 20% level) of a localized spectro-temporal mask, defined as a Gaussian envelope fit to each filter (see the main text and the Methods).

We consider two instantiations of the proposed framework. The *Feature-Based Model* operates directly in the time-frequency domain and operates linearly without constraints on the STRFs. We provide relevant theoretical results and validate the model on behavioral tasks for which physiological results are available, demonstrating that the resulting STRF adaptation patterns directly reflect task-relevant acoustic features. Next, we generalize the framework by considering an *Object-Based Model* that operates on the spectro-temporal modulation profiles of the STRFs with specific constraints on the magnitude and phase of the STRFs. By acting on an abstracted representation of the raw acoustics, this model therefore reflects a form of object-based attention. Again, we present theoretical results for this model. Predictions for behavioral tasks that could be readily evaluated in neurophysiological studies are also provided.

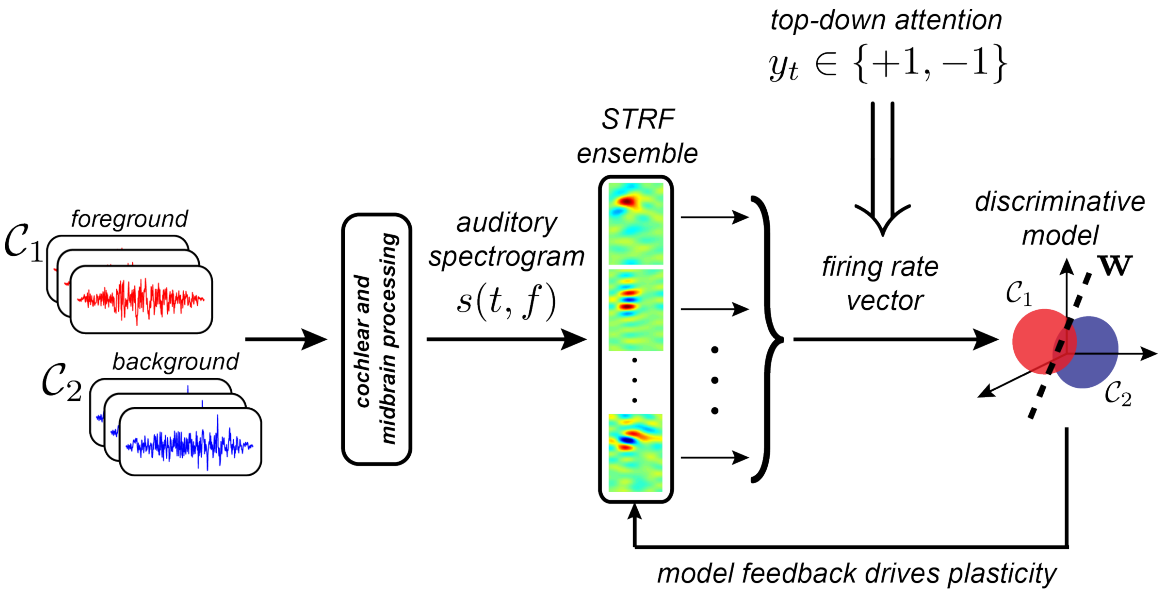

Figure 3.2: Proposed discriminative framework for attention-driven plasticity. Examples of foreground and background stimuli are passed through a model of the auditory periphery, and the resulting auditory spectrogram is analyzed by a bank of STRFs derived from recordings from ferret primary auditory cortex. Top-down attention acts to assign a behaviorally meaningful categorical label to observed population responses, which are subsequently discriminated using logistic regression. Feedback from the discriminative model, in the form of the regressor prediction error, iteratively adapts the shapes of the STRFs to improve prediction of foreground vs. background sounds.

3.3.3 Feature-Based Model: theoretical results

In the time-frequency domain, we model neural firing rate as

$$r_k(t) = \sum_f (m_k(t, f) \cdot h_k^A(t, f)) *_t s(t, f) \tag{3.2}$$

where $h_k^A(t, f) \in \mathbb{R}^{F \times T}$ denotes an STRF we seek to adapt, $*_t$ denotes convolution in time, $m_k(t, f) \in [0, 1]$ is a Gaussian-shaped spectro-temporal mask, and $s(t, f)$ is the stimulus spectrogram. The mask models physiological bounds on synaptic input and temporal integration that are typically observed in auditory cortical neurons. Later, we will observe that it guarantees that induced STRF adaptations are also spectro-temporally local. The mask is automatically determined by performing a least-squares fit of a Gaussian envelope to a rectified STRF (see Methods), and ellipses illustrating the coverage of the masks are shown in Fig. 3.1. Finally, let $\mathbf{r}_t = [1, r_1(t), r_2(t), \cdots, r_K(t)] \in \mathbb{R}^{K+1}$

denote an augmented ensemble response.

We model the influence of the top-down attentional signal $\mathcal{A}_t$ as the assignment of a behaviorally relevant categorical label $y_t \in \{+1, -1\}$ to an observed ensemble response $\mathbf{r}_t$, where $y_t = +1$ is associated with a target class of stimuli and $y_t = -1$ is associated with a reference class. To improve discrimination between target and reference stimuli, we assume that attention acts to vary the shapes of the STRFs in order to maximize the conditional likelihood of the labels. A simple model to quantify this notion is logistic regression, where we model the conditional likelihood as

$$p(Y_t = y_t | \mathbf{r}_t, \mathbf{w}) := \sigma(y_t \mathbf{w}^T \mathbf{r}_t) \tag{3.3}$$

where $\sigma(\alpha) = 1/(1 + \exp(-\alpha))^{-1}$ is the logistic function and $\mathbf{w} = [w_0, w_1, \cdots, w_K] \in \mathbb{R}^{K+1}$ is a vector of regression coefficients [133].

To induce task-driven changes in the STRFs, we define the following objective function:

$$J(\mathbf{w}, \mathcal{H}_A) := \underbrace{\frac{1}{2}||\mathbf{w}||_2^2 - C \cdot \left\langle \log \sigma \left(y_t \mathbf{w}^T \mathbf{r}_t \right) \right\rangle_t}_{\text{Discriminability}} + \underbrace{\frac{\lambda}{2} \sum_k ||h_k^0(t,f) - h_k^A(t,f)||_F^2}_{\text{Stability}} \tag{3.4}$$

The discriminability terms correspond to the average conditional log-likelihood of the attentional labels with l_2 regularization to prevent the regression coefficients from growing too large and overfitting available training stimuli. The stability term corresponds to an l_2 regularizer on the adapted STRF coefficients that controls "how far" the adapted STRFs can vary from their original versions. This reflects the idea that STRFs resist change and seek to return to their nominal shape upon task completion [78]. Finally, the balance between discriminability vs. stability is controlled by choice of hyperparameters (C, λ).

We optimized $J(\mathbf{w}, \mathcal{H}_A)$ using block coordinate descent, alternating between two minimization problems:

$$\arg\min_{\mathbf{w}} J(\mathbf{w}, \mathcal{H}_A) \;\; \text{subject to} \;\; w_k \geq 0, \quad k = 1, 2, \cdots, K \tag{P1}$$

$$\arg\min_{\mathcal{H}_A} J(\mathbf{w}, \mathcal{H}_A) \tag{P2}$$

We will show below that non-negativity constraints on the regression coefficients are necessary for encoding task valence during adaptation.

Because $J(\mathbf{w}, \mathcal{H}_A)$ is a sum of convex functions, and the constraints on (P1) are convex, each subproblem is therefore convex with a unique global minimum. Furthermore, since each update to $\mathbf{w}$ and $\mathcal{H}_A$ does not increase the value of $J(\mathbf{w}, \mathcal{H}_A)$, alternating updates to $\mathbf{w}$ and $\mathcal{H}_A$ guarantee convergence to a local minimum of the overall objective function [134, 135]. Intuition for this result can be gained by examining the sequence

$$J(\mathbf{w}^{(0)}, \mathcal{H}_A^{(0)}) \geq J(\mathbf{w}^{(1)}, \mathcal{H}_A^{(0)}) \geq J(\mathbf{w}^{(1)}, \mathcal{H}_A^{(1)}) \geq \cdots \geq J(\mathbf{w}^{(j+1)}, \mathcal{H}_A^{(j)}) \geq J(\mathbf{w}^{(j+1)}, \mathcal{H}_A^{(j+1)}) \geq \cdots$$

The solutions to both (P1) and (P2) are found numerically by searching for stationary points of the respective objective functions, i.e., when $\nabla_{\mathbf{w}} J(\mathbf{w}, \mathcal{H}_A) = 0$ and $\nabla_{h_k^A(t,f)} J(\mathbf{w}, \mathcal{H}_A) = 0$. For the regression coefficients, upon convergence of (P1), and assuming the minimum lies within the feasible set formed by the constraints on the w_k, the regression coefficient vector can be written as

$$\mathbf{w} = C \left\langle y_t \cdot [1 - \sigma(y_t \mathbf{w}^T \mathbf{r}_t)] \cdot \mathbf{r}_t \right\rangle_t \tag{3.5}$$

We interpret the term $[1 - \sigma(y_t \mathbf{w}^T \mathbf{r}_t)]$ as a "prediction error" and consequently hard-to-predict responses have more influence on choice of the optimal regression coefficients. Moreover, because the w_k for $k > 0$ are constrained to be nonnegative, those coefficients can be thought of as a *population gain vector* that applies more weight to task-relevant vs. task-irrelevant neurons.

Next, upon convergence of (P2), the adapted STRFs are found as

$$h_k^A(t, f) = h_k^0(t, f) + \frac{C}{\lambda} \cdot w_k \cdot m_k(t, f) \left\langle y_{t'} \cdot [1 - \sigma(y_{t'} \mathbf{w}^T \mathbf{r}_{t'})] \cdot s(t' - t, f) \right\rangle_{t'} \tag{3.6}$$

Eq. 3.6 contains the main theoretical result of the Feature-Based Model and shows how STRF plasticity predicted by the proposed framework is consistent with the contrast filtering hypothesis. First, attention-induced STRF plasticity directly reflects the spectro-temporal structure and features of the (time-reversed) target and reference stimuli, as given in the averaging term. The impact of the stimulus on adaptation at each time is proportional to the difficulty of predicting its corresponding label. Second, because we have constrained the regression coefficients w_k for $k > 0$ to be non-negative, the behavioral meaning of the labels is preserved so that acoustic features of the target ($y_t = +1$) are guaranteed to be *enhanced* whereas those of the reference ($y_t = -1$) are *suppressed*. Third, STRF plasticity is guaranteed to be local as a consequence of multiplying the sum with the Gaussian-shaped spectro-temporal mask $m_k(t, f)$. Finally, the first term encourages stability in the STRFs by resisting change from their original shapes, the magnitude of the effect being controlled by C and λ.

3.3.4 Feature-Based Model: validation

We validate the model by simulating task-driven plasticity on a number of spectral behavioral tasks that have been explored in studies of auditory cortex. We first consider a *tone detection* task, where an animal is trained to detect an isolated tone in the context of a broadband noise reference [16]. The second is a *chord detection* task, where an animal is trained to detect a multi-tone complex in the context of a broadband noise reference [75]. Finally, we consider a *tone discrimination* task, where an animal is trained to detect a target tone in the context of a specified reference tone [17]. The details of the stimuli used for each task are provided in Table 3.1 and the details of stimulus construction are provided in the Methods.

Table 3.1: Details of the tasks considered for the Feature-Based Model.

Task	Target	Reference
Tone detection	0.25, 0.5, 1, 2, 3.25 kHz	TORCs
Chord detection	0.25/0.5/0.75, 0.5/0.75/2, 0.5/1/2, 1/2/1.5, 1.75/2/3.25 kHz	TORCs
Tone discrimination	0.25, 0.25, 0.5, 0.5, 1 kHz	0.5, 1, 1, 2, 2 kHz

To visualize the effects of attention on the shapes of the receptive fields, we consider the difference between the Euclidean-normalized active and passive STRFs ($\Delta STRF$); examples of the induced adaptation patterns for the spectral tasks are shown in Fig. 3.3. For tone detection, panels A and B illustrate that target tones (red arrows) induce local, excitatory changes in the STRFs at the target frequencies. This is apparent from the active STRFs (middle subpanels) as well as from the difference STRF (right subpanels). The difference STRF also reveals that the effect of the noise reference is to introduce a small degree of suppression within the mask and surrounding the tone. Similar effects are observed for the chord detection in panels (C and D): target tones induce local, excitatory changes, with suppression around and in between the target tones. Finally, shown in panels E and F are example adaptation patterns for the tone discrimination task. We observe that target tones induce excitatory changes whereas reference tones (blue arrows) induce inhibitory changes in the active STRFs.

We quantify population effects using approaches described in previous physiological studies (see, e.g., [16]), and the results are summarized in Fig. 3.4. First, to visualize population effects across a number of targets (references), we compute $\Delta STRF$ aligned at the target (reference) frequencies, and average across all ensembles and target (reference) tones. Next, in order to quantify the size of the attentional effect, we compute the relative change of STRF gain, at the location of maximum difference in the target (reference) channel, between the passive and active settings; we refer to this as ΔA and subscript accordingly for each task.

For tone detection, panel A shows that across all targets and ensembles, active attention simulated by the model induces local, excitatory changes in the STRFs at the target tone, with inhibitory changes spectrally adjacent to the target. Panel B shows that the distribution of ΔA_{TGT} is overwhelmingly excitatory (mean $= +50.87 \pm 6.7\%$ s.e.m.) with a heavy tail to the right. For each ensemble and across all targets, excitatory changes are significant ($p \ll 0.001$, t-test and Wilcoxon signed-rank test). Importantly, similar observations have been made in ferret recordings by Fritz et al. [16].

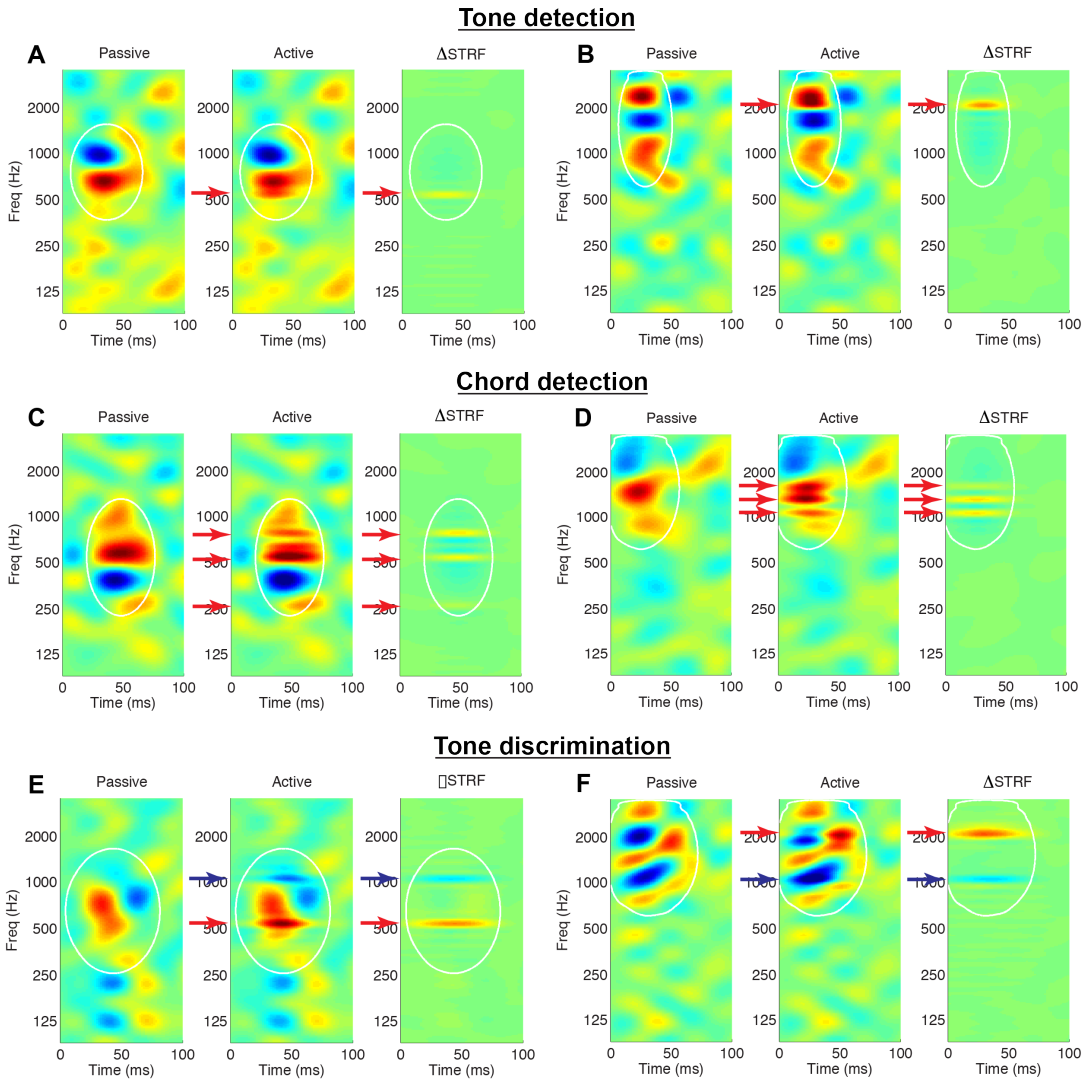

Figure 3.3: Validation of the Feature-Based Model on a variety of behavioral tasks. Each panel shows an STRF in the passive and active behavioral state, and the difference STRF illustrates the effects of the model on STRF shape. (A, B) *Tone detection:* Target tones (red arrows) elicit increased excitation at the target frequency. The difference pattern also reveals a small degree of inhibition at non-target frequencies within the mask. (C, D) *Chord detection:* Target tones elicit increased excitation at each of the frequencies in the target complex, with regions of suppression between and outside the targets. (E, F) *Tone discrimination:* Target tones elicit increased excitation whereas reference tones (blue arrows) are suppressed. White lines: isoline contours of the spectro-temporal mask at the 20% level. STRFs are interpolated for display. Examples shown for $\lambda = 10^{-4.5}, C = 10^{-3}$.

For chord detection, the target stimuli comprise three tones, some of which may be near or far to a given neuron's best frequency (BF). Consequently, we expect that tones near BF would induce stronger plasticity effects compared to those far from BF. We verify this by computing the average $\Delta STRF$ aligned to target tones nearest to and furthest from BF, and these results are shown in panel C. As shown, tones near BF induce stronger local excitatory changes compared to tones far from BF. As suggested previously in Fig. 3.3, the suppressed sidebands surrounding the target tones show that the active STRFs were suppressed in between each of the target tones. The inhibitory effect is also relatively stronger for tones near BF compared to those far from BF. Importantly, this analysis has parallels with that of [75], and we again find a general correspondence with those previously reported results. Finally, in panel D, we consider the distribution of ΔA_{TGT} for near vs. far targets across all ensembles. These distributions show that changes at the target tones are overwhelmingly excitatory (mean $+48.4 \pm 9.8\%$ vs. $+35.1 \pm 7.4\%$, near vs. far, s.e.m.) with heavy tails to the right, and are stronger for targets near BF vs. those far from BF. For each ensemble and across all targets, excitatory changes are significant ($p < 0.03$, t-test and Wilcoxon signed-rank test).

Next, for tone discrimination, we considered $\Delta STRF$ aligned to both the reference and target tones; these results are shown in panel E averaged across all ensembles and target/reference combinations. As shown, the model induces local, inhibitory changes at the reference compared to local, excitatory changes at the target. Importantly, these differential plasticity effects are consistent with observations by Fritz et al. from a ferret study [17]. On the left side of panel F, we show the distribution of ΔA at the target and reference tones. As predicted by the model, attention induces excitatory changes at the target (red, mean $= +66.0 \pm 12.4\%$ s.e.m.) while changes at the reference are inhibitory (blue, mean $= -71.4 \pm 5.0\%$ s.e.m.). For each ensemble and across all tasks, excitatory and inhibitory changes are significant ($p \ll 0.001$, t-test and Wilcoxon signed-rank test).

Finally, we verify that non-negativity constraints imposed on the regression coefficients allow the model to capture the behavioral meaning associated with the target and reference stimuli.

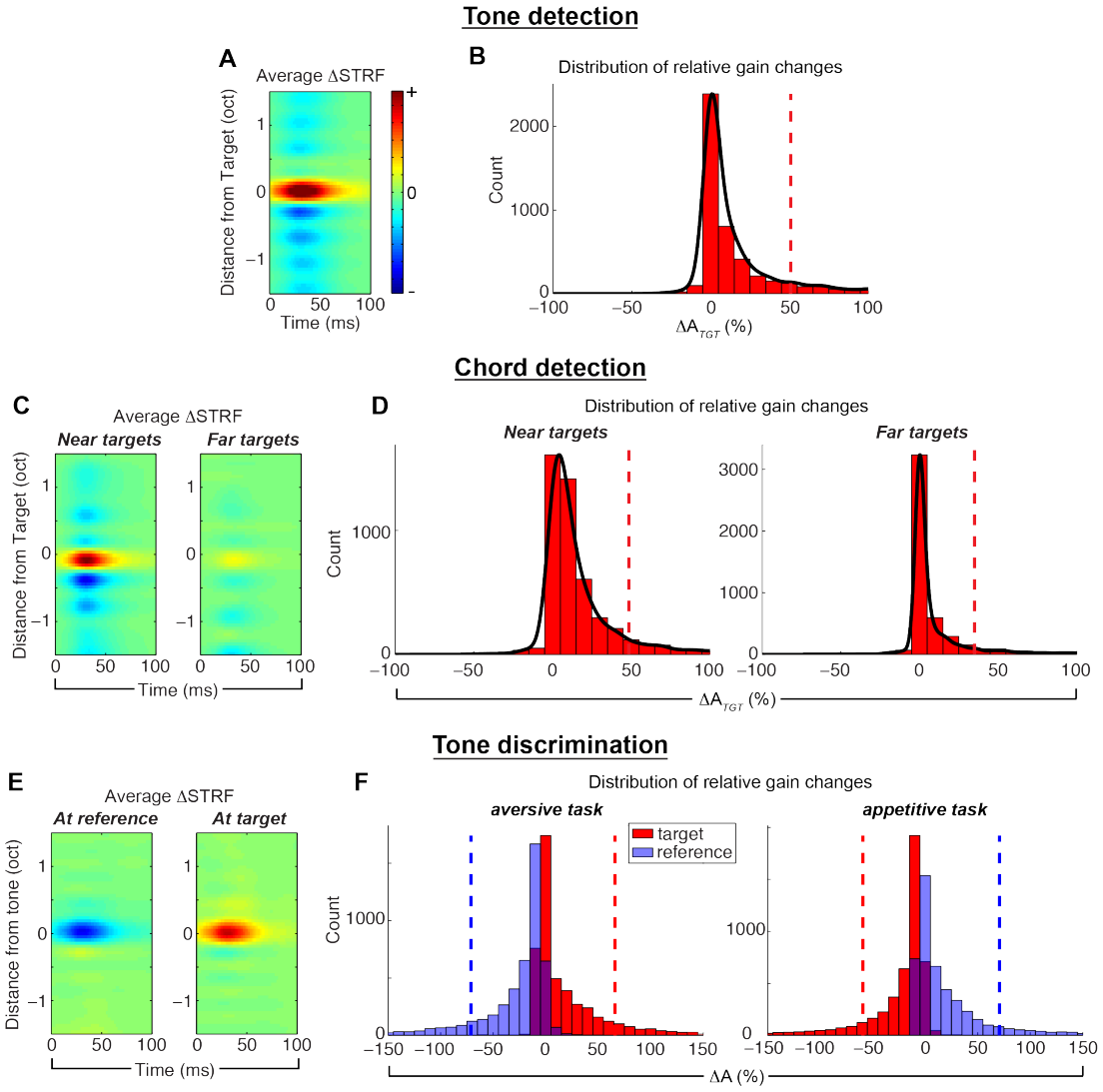

Figure 3.4: Population analysis of the Feature-Based Model. *Tone detection:* The average $\Delta STRF$ in panel A, computed by aligning all difference STRFs at the target frequency, shows that target tones elicit increased excitation, whereas the broadband noise reference induces suppression in areas spectrally adjacent to the target. Panel B shows that relative gain changes at the target due to attention are overwhelmingly excitatory. *Chord detection:* Panel C shows the average $\Delta STRF$ aligned to targets nearest (left) and farthest (right) from a neuron's BF. Targets close to BF induce much larger excitatory changes than those farthest away, and this pattern is also observed in ΔA_{TGT} in panel D. Suppressive effects are similar to those observed in single tone detection tasks. *Tone discrimination:* Panel E shows the average $\Delta STRF$ aligned at the reference and target tones in an aversive task setup. STRF changes are suppressive at the reference and excitatory at the target, which is also observed in patterns of ΔA (panel F, left). However, when the behavioral meaning of the target and reference is reversed, as in an appetitive task, STRF plasticity patterns are similarly reversed (panel F, right). Average $\Delta STRF$ patterns are interpolated for display. Dashed vertical lines denote population means. Results shown for $\lambda = 10^{-4.5}, C = 10^{-3}$.

As demonstrated in a ferret study [76], differences in animal training for aversive tasks (target associated with negative reward) vs. appetitive tasks (target associated with positive reward) yield excitatory and inhibitory patterns at the target and reference tones that are flipped versions of each other. In our framework, this is achieved by simply flipping the sign of the labels associated with the target and reference stimuli. The recomputed ΔA distributions after flipping labels are shown on the right side of panel F. As shown, the distributions of ΔA for the appetitive task are flipped versions of the aversive task (target mean = $-59.1 \pm 8.5\%$, reference mean = = $+70.5 \pm 7.9\%$, s.e.m.). For each ensemble and across all tasks, excitatory and inhibitory changes are significant ($p \ll 0.001$, t-test and Wilcoxon signed-rank test). These results confirm that the model does indeed capture aspects of task structure.

3.3.5 Object-Based Model: theoretical results

The Feature-Based Model, while sufficient to account for adaptation patterns in purely spectral tasks, is restricted to act at the level of the raw spectro-temporal features that characterize the task-related stimuli (see Eq. 3.6). However, accumulating evidence suggests that top-down attention can instead modulate neural representations at the level of auditory objects [6, 7, 54, 136, 137]. Broadly speaking, object-based attention refers to the selective allocation of cognitive resources to an *abstracted* representation of a stimulus. For our purposes, we interpret this as attention directed towards collections of features that may be used to distinguish broad stimulus classes from one another (e.g., speech vs. non-speech sounds). One way to abstract acoustic information in a spectrogram is to consider its representation in the Fourier domain, where the strength of observed spectro-temporal modulations (i.e., the Fourier magnitude profile) could be considered separately from the relative activation of the modulations to one another (i.e., the Fourier phase profile). Thus, attention directed towards a collection of spectro-temporal dynamics, rather than the relative timings of the observed acoustics, represents an instantiation of object-driven attention. For example, in complex acoustic scenes, a listener may wish to attend to conspecific vocalizations in noisy natural

environments, retuning cognitive resources to enhance responses to time-varying harmonicity cues (which are often characteristic of animal communication sounds) while suppressing those to the din of spectro-temporally broad background interference.

Furthermore, there is neurophysiological evidence suggesting that receptive field plasticity that reflects differences in stimulus modulation profiles contributes to improved performance of behavioral tasks. For example, [138] showed that the temporal modulation profiles of A1 neurons in monkeys trained to discriminate temporally modulated tone sequences adapted to enhance responses of faster target modulations (associated with a negative reward) while suppressing responses to slower reference modulations. In a visual study, [50] found that the modulation profiles of spatio-temporal receptive fields in higher visual cortex adapted to match those of a target stimulus in both discrimination and search tasks. Finally, [77] recently demonstrated that the joint spectro-temporal modulation profiles of STRFs in ferret A1 adapted to reflect the difference in modulation characteristics of upward vs. downward moving tone pips. Motivated by these examples, we sought to extend the proposed framework to circumstances where task-relevant stimuli could be discriminated based on differences in their spectro-temporal dynamics, and we directly modified STRF shapes in the Fourier domain accordingly.

We begin by first modifying the firing rate model as

$$r_k(t, f; m) = h_k^A(t, f) *_{tf} s_m(t, f) \tag{3.7}$$

with corresponding *modulation domain* representation

$$|R_k(\omega, \Omega; m)| = |H_k^A(\omega, \Omega)| \cdot |S_m(\omega, \Omega)| \tag{3.8}$$

where $*_{tf}$ denotes convolution in time and frequency, and $R_k(\omega, \Omega; m)$, $H_k^A(\omega, \Omega)$, and $S_m(\omega, \Omega)$ are the 2D Discrete Fourier Transforms of firing rate, STRF, and the m'th stimulus token, respectively. In the modulation domain, ω characterizes modulations along the temporal axis (*rate*, in Hz) whereas

Ω characterizes modulations along the spectral axis (*scale*, in cycles/octave). For technical reasons (see Supp. Text 1 in Appendix B), in this instantiation of the model we forego use of the mask, but we address its absence later in the Discussion.

The development of the Object-Based Model development mirrors that of the Feature-Based Model. First, we form a firing rate vector as $\mathbf{R}_m = [1, \sum_{\omega\Omega} |R_1(\omega, \Omega)|, \cdots, \sum_{\omega\Omega} |R_K(\omega, \Omega)|] \in \mathbb{R}^{K+1}$. Next, we again use logistic regression and Euclidean norm to quantify the balance between discriminability and stability, and define the objective function

$$J(\mathbf{w}, \widehat{\mathcal{H}}_A) : \underbrace{\frac{1}{2}||\mathbf{w}||_2^2 - C \cdot \left\langle \log \sigma \left(y_m \mathbf{w}^T \mathbf{R}_m \right) \right\rangle_m}_{\text{Discriminability}} + \underbrace{\frac{\lambda}{2} \sum_k ||\Delta_k||_F^2}_{\text{Stability}} \tag{3.9}$$

where $\mathbf{w}$ is defined as before, $\widehat{\mathcal{H}}_A := \{|H_k^A(\omega, \Omega)|\}_{k=1}^K$, and $\Delta_k := |H_k^0(\omega, \Omega)| - |H_k^A(\omega, \Omega)|$.

To optimize Eq. 3.9, we again used block-coordinate descent, alternating between solving two convex subproblems:

$$\arg\min_{\mathbf{w}} \; J(\mathbf{w}, \widehat{\mathcal{H}}_A) \;\; \text{subject to} \;\; w_k \geq 0, \quad k = 1, 2, \cdots, K \tag{P3}$$

$$\arg\min_{\widehat{\mathcal{H}}_A} \; J(\mathbf{w}, \widehat{\mathcal{H}}_A) \;\; \text{subject to} \;\; |H_k^A(\omega, \Omega)| \geq 0, \quad \forall\, k, \omega, \Omega \tag{P4}$$

The constraints on (P4) are required since modulation profiles $|H_k^A(\omega, \Omega)|$ are necessarily nonnegative.

Optimizing (P3) yields regression coefficients similar to those in Eq. 3.5. Next, upon convergence of (P4), and assuming the minimum lies within the feasible set formed by the constraints on $|H_k^A(\omega, \Omega)|$, the adapted STRF modulation profiles can be written as

$$|H_k^A(\omega, \Omega)| = |H_k^0(\omega, \Omega)| + \frac{C}{\lambda} \cdot w_k \cdot \left\langle y_m [1 - \sigma(y_m \mathbf{w}^T \mathbf{R}_m)] \cdot |S_m(\omega, \Omega)| \right\rangle_m \tag{3.10}$$

Eq. 3.10 contains the main theoretical result of the Object-Based Model, which is again

consistent with the contrast filtering hypothesis and similar in spirit to the Feature-Based Model. First, attention-induced STRF plasticity directly reflects the spectro-temporal modulation profiles of the target and reference stimuli, as given in the averaging term. The impact of each stimulus sample on adaptation is proportional to the difficulty of predicting its corresponding label. Again, because we have constrained the regression coefficients w_k to be non-negative, the behavioral meaning of the labels is preserved so that acoustic features of the target are *enhanced* whereas acoustic features of the reference are *suppressed*. The first term acts to resist changes from the initial STRF modulation profile, the magnitude of the effect being controlled by C and λ. Finally, we note that to visualize the adapted STRFs in time-frequency, we use the original, unmodified phase of the passive STRF.

3.3.6 Object-Based Model: predictions

To evaluate the predictions of the Object-Based Model, we consider two behavioral tasks that can be readily explored in animal studies. The first is *spectro-temporal modulation noise discrimination*. Classes of natural stimuli often overlap in terms of their spectral and temporal modulation content but are distinguished by the additional presence or absence of energy at certain rates and scales, e.g., speech vs. speech+noise, conspecific vocalizations in noisy natural environments, etc. In this spirit, we synthesize complex spectro-temporal noise stimuli that share a broad range of modulations but are distinguished by additional energy at downward vs. upward rates and scales. The stimuli are generated by specifying the energy distribution of the target and reference modulation profiles, coupling them with random phase, and performing an inverse 2D Fourier transform to obtain the stimulus spectrograms. An example of this process is shown in Fig. 3.5A and B for what we term a *Broadband Down* (BB Down) target and *Broadband Up* (BB up) reference. The ellipses in panel A represent Gaussians in the modulation domain, and the dashed lines indicate the set of modulations that are shared between the target and the reference. Here the target is characterized by the addition of a range downward modulations centered at (+16 Hz, 0.25 cyc/oct) whereas the reference is a flipped version of the target, containing added upward modulations. After coupling

with random phase and performing an IDFT, we obtain the spectrograms shown in panel B. For this work we consider four discrimination tasks, the details of which are provided in the Methods and summarized in Table 3.2.

Table 3.2: Details of the tasks considered for the Object-Based Model.

Task		Target	Reference
Spectro-temporal modulation noise discrimination	*Narrowband Up* (NB up)	(-10 Hz, 1 cyc/oct)	(+10 Hz, 1 cyc/oct)
	Narrowband Down (NB down)	(+10 Hz, 1 cyc/oct)	(-10 Hz, 1 cyc/oct)
	Broadband Up (BB up)	(-16 Hz, 0.25 cyc/oct)	(+16 Hz, 0.25 cyc/oct)
	Broadband Down (BB down)	(+16 Hz, 0.25 cyc/oct)	(-16 Hz, 0.25 cyc/oct)
Click rate discrimination		18 Hz	5 Hz
		24 Hz	7 Hz
		32 Hz	9 Hz

The second task we consider is *click rate discrimination*, where the goal is to discriminate a fast from a slow click train. To the best of our knowledge, there have been no studies that report population STRF plasticity patterns for this task (though see the examples presented in [17]). We synthesize idealized click trains directly in the time-frequency domain, and an example is shown in Fig. 3.5C with its corresponding spectro-temporal modulation profile shown in panel D. By construction, the broadband clicks contain energy only at 0 cyc/oct, i.e., purely temporal modulations. Consequently, adaptation of the modulation profiles will only occur at this scale (panel D, circled), so we restrict our population analysis accordingly. Panel E shows the temporal modulation profile of an example stimulus, with a peak at 18 Hz and associated harmonics. For this work, we consider three click discrimination tasks, and the specific details of the stimuli are provided in the Methods and summarized in Table 3.2.

For spectro-temporal modulation noise discrimination, we find model-induced adaptation patterns for individual neurons that are consistent with the contrast filtering hypothesis. Shown in Fig 3.6A and B are two examples of the model effects when engaged in two modulation noise discrimination tasks. As before, the top rows of each panel show the passive, active, and normalized difference STRF. Here, however, the bottom rows show the passive, active, and difference MTFs (ΔMTF) over a broad range of rates and scales. In both examples, the model predicts STRF

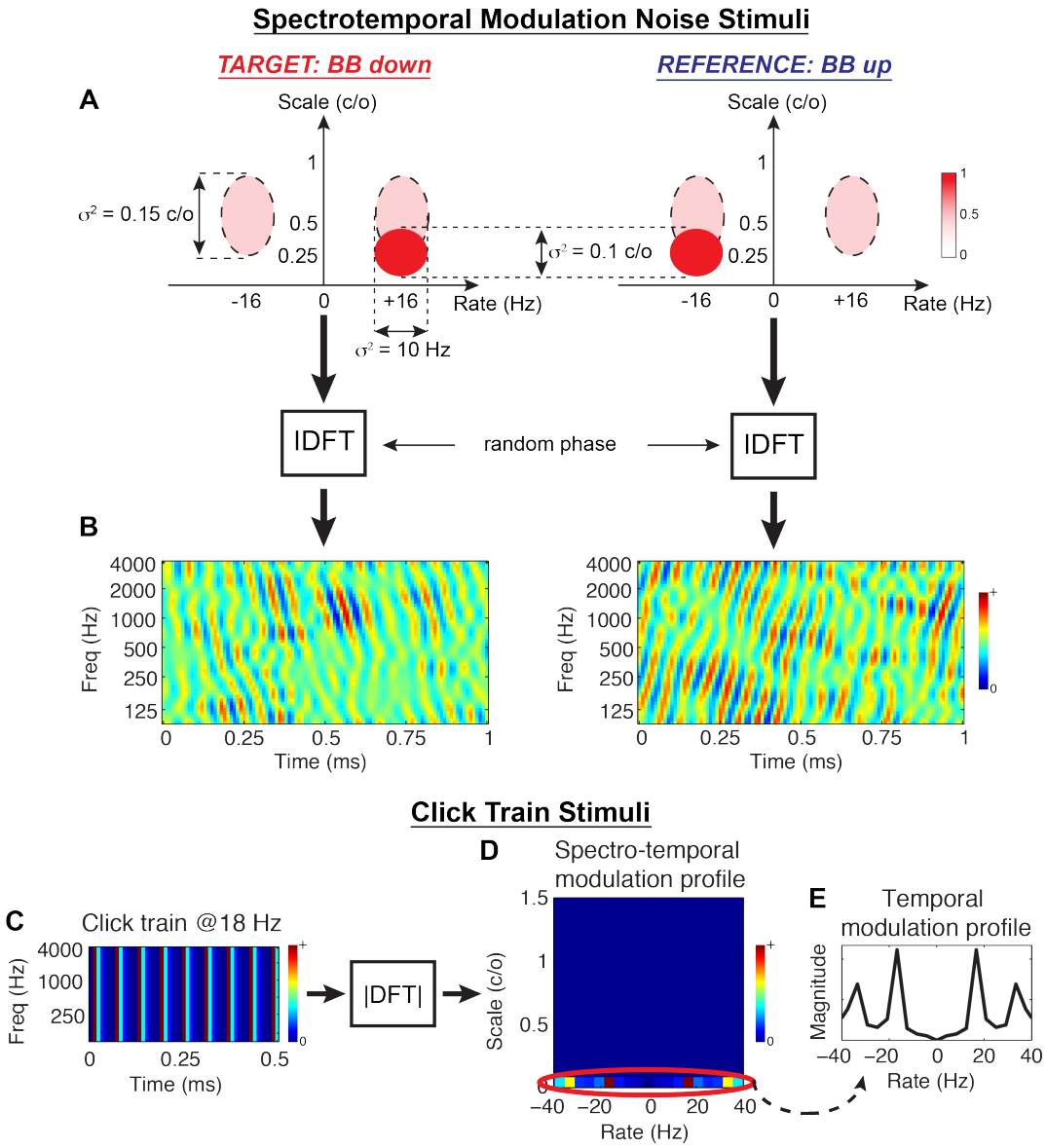

Figure 3.5: Stimulus design for testing the Object-Based Model. *Spectro-temporal modulation noise stimuli:* As illustrated in panel A, noise stimulus profiles are designed to overlap in the modulation domain over a broad range (dashed ellipses), and each class is distinguished by added energy centered at a prescribed rate and scale (solid ellipses). The modulation profiles are coupled with random phase, and an inverse 2D Discrete Fourier Transform is performed to synthesize the stimuli in time-frequency (panel B). In this example, a target stimuli characterized by broad, downward drifting modulations is contrasted with a reference of broad, upward drifting modulations. *Click train stimuli:* Simple broadband click trains are synthesized directly in time-frequency (panel C), and necessarily only have energy in the spectro-temporal modulation domain at 0 c/o (panel D). For analysis purposes, we consider changes in the STRF temporal modulation profiles only at this scale (panel E).

plasticity that reorients and sharpens tuning for target modulations (downward for panel A, upward for panel B). This effect is also clear from the difference MTFs, which show explicit enhancement of target and suppression of reference modulations.

We also find that population patterns of plasticity are broadly consistent with the contrast filtering hypothesis. We summarized these population patterns in the modulation domain by averaging ΔMTF across all neurons; these results are shown in panel C. For each task, we find that on average target modulations are enhanced whereas reference modulations are suppressed. We also consider model effects on the directional preference of the STRFs, as quantified by a directionality index (DIR, see Methods). In general, positive DIR indicates a preference for downward modulations whereas negative DIR indicates a preference for upward moving modulations. The effect of the model between the passive and active settings can be measured by computing the change in directionality, defined as $\Delta DIR := DIR_A - DIR_P$ (where the subscripts denote active and passive, respectively). Thus, positive values of ΔDIR indicate a shift towards a preference for downward modulations, whereas negative values of ΔDIR indicate a shift towards a preference for upward moving modulations. Panel D shows the distributions of ΔDIR for each tasks. As shown, upward moving targets induce a significant directional preference for upward modulations, and similarly so for downward moving targets ($p < 0.01$, Wilcoxon signed-rank test).

For click rate discrimination, we find that the Object-Based Model induces plasticity patterns in individual neurons that are consistent with the contrast filtering hypothesis, with effects that are evident in both the original time-frequency space as well as in the temporal modulation profiles. As expected, we find that modulations at the target click rate are enhanced, whereas modulations at the reference click rate are suppressed. Shown in Fig. 3.7A and B are two examples of the simulated plasticity effects for this task. The top row of each panel shows the passive, active, and normalized difference STRF ($\Delta STRF$) whereas the bottom row shows the passive, active and difference modulation transfer functions (ΔMTF) at 0 cyc/oct. For both examples, it is clear in the time-frequency domain that the model induces purely temporal adaptation, as evidenced by the

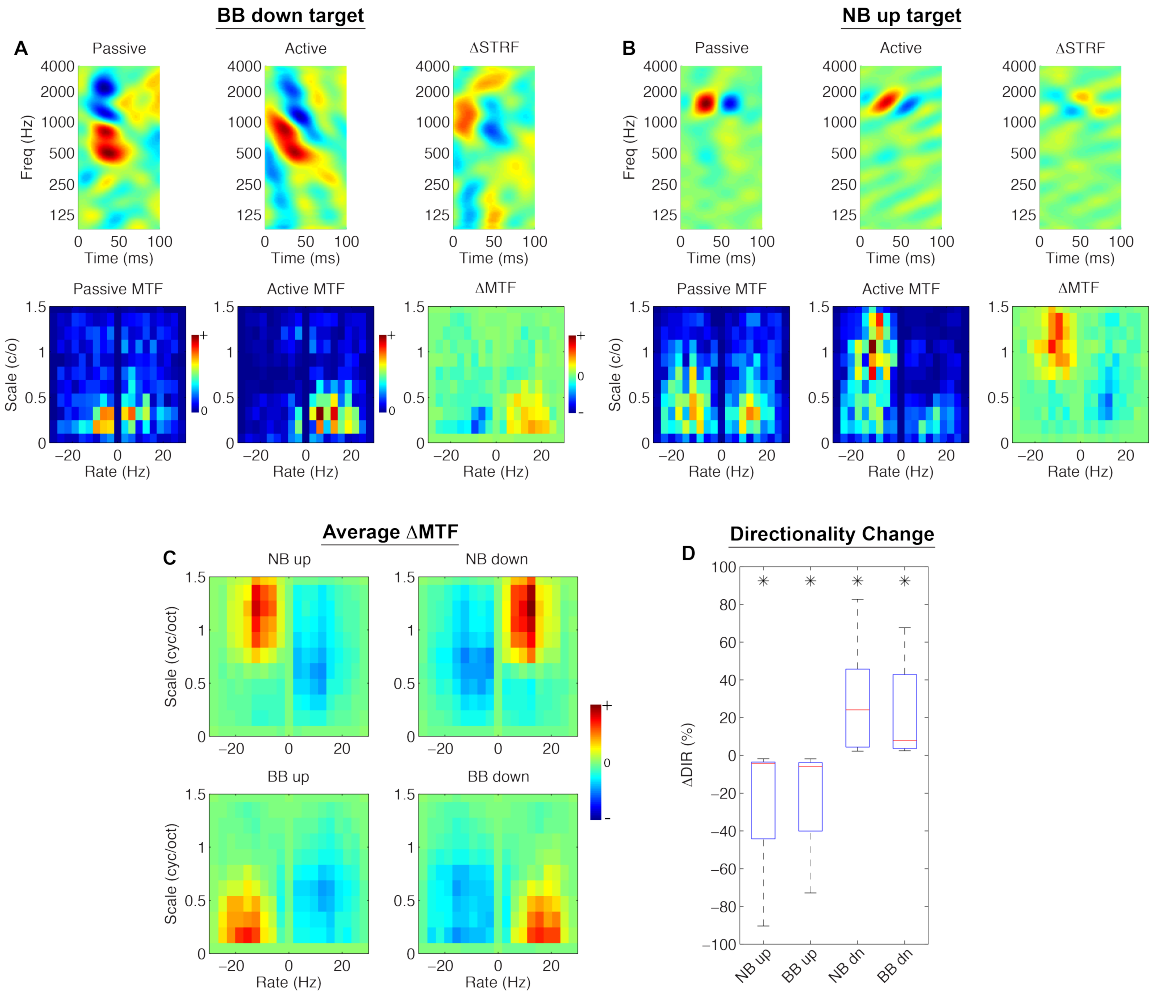

Figure 3.6: Object-Based Model predictions for spectro-temporal modulation noise discrimination. In panels A and B, the top row shows the passive, active, and difference STRF, whereas the bottom row shows the passive, active, and difference MTF (note: ΔMTF is *not* the modulation profile of $\Delta STRF$). The active STRFs are characterized by downward or upward changes in orientation depending on the target stimulus class (A and B, respectively). Furthermore, the difference MTF illustrates that target modulations are enhanced whereas reference modulations are attenuated. Panel C shows that across all tasks and populations, target modulations are enhanced and reference modulations are suppressed. Finally, panel D shows that changes in directional preference of the active STRFs, as quantified by ΔDIR, reflect a significantly increased sensitivity to the target class ($* : p < 0.01$, Wilcoxon signed-rank test). Results shown for $\lambda = 10^{-4}, C = 0.5$.

vertical bars in $\Delta STRF$. These changes had an apparent effect on the temporal bandwidth of the main excitatory subfield of the active STRFs, in some cases inducing a narrowing and in others a broadening of the subfields (A and B, respectively). Furthermore, in the modulation domain, it is clear from the difference MTFs that energy at the target rate is enhanced whereas energy at the

reference click rate is suppressed.

We again find that population patterns of plasticity are broadly consistent with the contrast filtering hypothesis, with adapted neurons exhibiting increased (decreased) sensitivity to the target (reference) click rates. To summarize these population patterns in the modulation domain, for each task we first averaged ΔMTF across all neurons and for clarity fold the modulation profile about 0 Hz rate; these results are shown in panel C. As shown, for each task ΔMTF is positive at the target click rate (and its harmonics) and negative at the reference click rate. Changes at the target and reference click rates are significant for each task ($p \approx 0$, Wilcoxon signed rank test). In panel D we also show the distribution of changes in the temporal bandwidth of the main excitatory subfields (ΔBW_T, see Methods). Here, negative values indicate temporal narrowing whereas positive values indicating temporal broadening. Temporal bandwidth in the active STRFs tends to be slightly, but significantly, narrowed (mean $\Delta BW_T = -0.53$ ms, $p < 0.01$, t-test). However, the distribution shows that while the changes are generally are quite subtle, a large number of neurons (40.4% across all tasks and ensembles) have an absolute change greater than 1 ms. Interestingly, the model predicts that excitatory subfields will both contract and expand as needed to enhance sensitivity to target click modulations, as indicated by both negative and positive values of ΔBW_T. Similar behaviors have been observed in neurophysiological studies yet to be published [139], though an exact quantification of this effect in experimental findings is not yet readily available.

3.4 Discussion

In this chapter, we have proposed and explored a discriminative framework for modeling task-driven plasticity in auditory cortical receptive fields. The framework predicts STRF adaptation patterns that are consistent with the contrast filtering hypothesis: that neural tuning characteristics at the level of primary auditory cortex adapt to enhance acoustic features of the foreground while actively suppressing those of the background. We proposed two instantiations of the framework: a

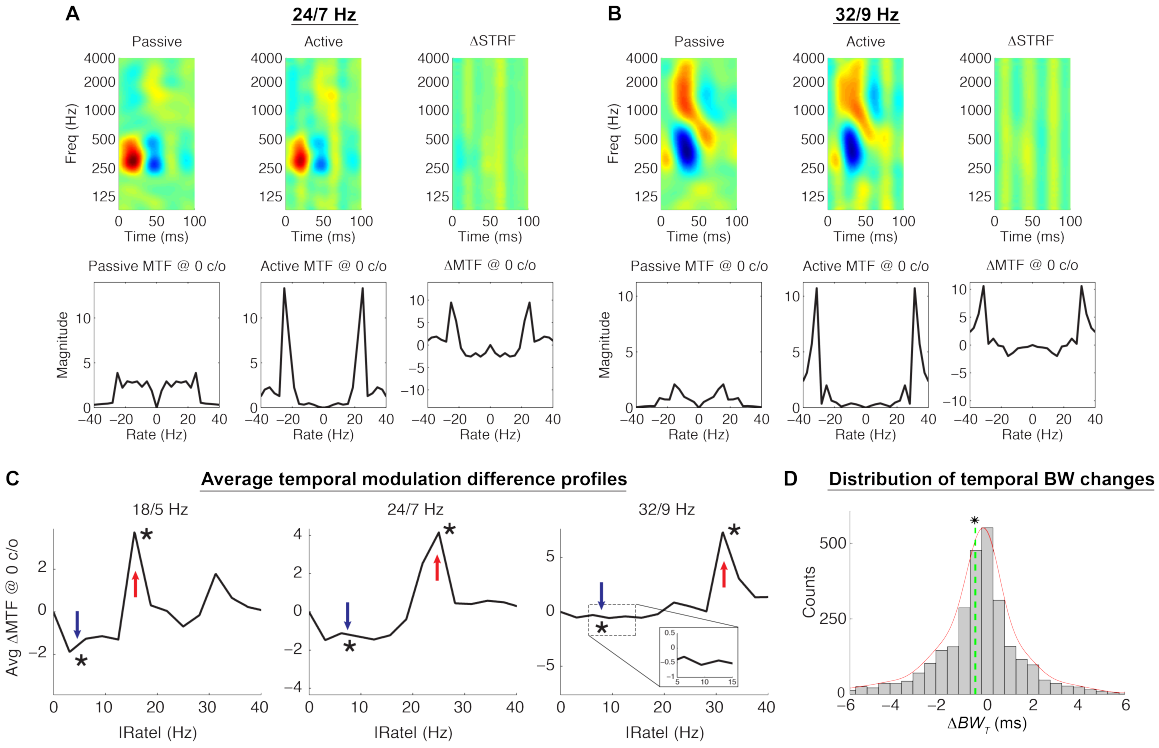

Figure 3.7: Object-Based Model predictions for click rate discrimination. In panels A and B, the top row shows the passive, active, and difference STRF, whereas the bottom row shows the passive, active, and difference MTF (note: ΔMTF is *not* the temporal modulation profile of $\Delta STRF$). The active STRFs are characterized by the addition of broadband temporal ripples that have the effect of slightly narrowing or broadening the main excitatory subfields of the STRFs (A and B, respectively). The difference MTFs at 0 c/o show enhancement and suppression of the target and reference click rates, respectively. Panel C shows the average ΔMTF at 0 c/o, folded at 0 Hz for clarity, for each of the click rate discrimination tasks, showing significant enhancement and suppression at the target and reference click rates, respectively ($* : p \approx 0$, Wilcoxon signed-rank test). Panel D shows the distribution of changes in temporal bandwidth for STRF main excitatory subfields across all tasks and ensembles. Temporal bandwidth, while on average slightly decreased (mean $\Delta BW_T = -0.53$ ms, dashed vertical line; $* : p < 0.01$, t-test), can be both increased and decreased by adaptation of the temporal modulation profile. Results shown for $\lambda = 10^{-4}, C = 0.5$.

Feature-Based Model that acts directly based on raw acoustic features in the time-frequency domain; and an Object-Based Model that acts in a stimulus phase-invariant fashion on an abstracted representation of the stimuli in the spectro-temporal modulation domain. We showed, via simulations of a number of spectral behavioral tasks, that the Feature-Based Model induced localized STRF adaptation that enhanced representation for the target tone while inducing mild sideband suppression (for tone/chord detection tasks) or narrowband suppression at the reference tone (for tone discrim-

ination tasks). Importantly, these results are consistent with plasticity patterns previously reported in neurophysiological studies [16, 17, 75]. We also showed, via the tone discrimination tasks, that switching the behavioral meaning associated with target and reference stimuli (i.e., by switching the model labels) induces opposite plasticity patterns. This is akin to modifying animal training protocol from an aversive to appetitive task structure where similar flipped adaptation patterns have been observed in ferret A1 neurons [76]. This suggests that the model captures aspects of task structure, which has yet to be explicitly accounted for by previous computational models [76, 80].

Next, we explored predictions of the Object-Based Model on tasks that could be readily evaluated in neurophysiological studies. We first considered the task of spectro-temporal modulation noise discrimination. This was intended to model naturalistic scenarios where a listener seeks to direct attention among acoustic classes of similar timbres, i.e., those that share a broad range of spectro-temporal modulations but differ based on the presence or absence of energy at a smaller set of rates and scales. For these stimuli, the model predicted enhancement at the subset of modulations that defined the target class, whereas we observed suppression at the subset of modulations that defined the reference class. The overall effects in time-frequency were an effective reorientation and sharpening of the STRFs to the target modulations, and we quantified these changes using a directionality measure that characterized a neuron's preference for downward vs. upward drifting modulations.

Finally, we considered the task of click rate discrimination for which, to the best of our knowledge, population patterns of STRF plasticity have yet to be reported (save for examples reported by [140]). The model predicted that for purely temporal tasks, the temporal modulation profile of the active STRFs is enhanced at the target click rate and suppressed at the reference click rate. This had the effect of introducing broadband, temporal ripples in time-frequency, as evidenced by the difference STRFs in Fig. 3.7A and B. While it has previously been observed in other animal models and temporal tasks that the temporal dynamics of cortical neurons can shift to become more responsive (i.e., reduced temporal bandwidth or latency) [140–142], the Object-Based

Model predicts that the main excitatory subfields of neurons can become either temporally narrower or broader so long as the overall temporal modulation profile is suitably adapted at the target and reference click rates.

3.4.1 An integrated framework for modeling attention-driven plasticity

Optimization within the proposed framework is by necessity constrained and iterative, due the need to alternate between solving two convex subproblems to determine optimal regression coefficients and STRF parameters. However, this approach may reflect an analogous iterative adaptation strategy among neural circuits in the cortical hierarchy thought to be involved in task-driven auditory attention. In particular, it has been suggested that attention involves an iterative circuit among basal forebrain, prefrontal cortex (PFC), and sensory cortex that, from a computational perspective, has a number of parallels with our proposed framework [140, 143, 144].

A simplified schematic of this process is shown in Fig. 3.8 and the basic process is as follows. Input acoustic stimuli are processed by a bank of cortical receptive fields, which project to, and receive projections from, executive control networks in PFC. Importantly, projections to PFC are gated according to behavioral and task salience, i.e., only task-relevant signals are passed along and processed [79]. Decoded signals in PFC in turn cause motor responses that prompt the listener to act (e.g., cease licking water in response to a target tone), and consequently induce plasticity to improve performance of the task. These feedback circuits likely involve nucleus basalis (NB) and the ventral tegmental area (VTA), basal forebrain areas that have been implicated in cortical plasticity [142, 145].

As annotated in the figure, we propose that the framework described in this study has useful parallels with the circuits enclosed within the dashed boxes and that, in general, the alternating optimization procedure reflects a biologically plausible strategy for fine-tuning sensory input based on task-performance. In particular, we argue that during active attention, the computational goal of top-down executive control circuits, like those in PFC, is to assign behaviorally meaningful categorial

labels to observed ensemble responses in primary auditory cortex (box A). Subsequent classification decisions in turn induce appropriate motor responses to perform the task at hand. Furthermore, as seen in Eqs. 3.6 and 3.10, the magnitude of plasticity effects in the model is directly proportional to the magnitude of regressor prediction errors (box B). This has parallels with behavioral results in ferret studies, where the magnitude of STRF plasticity effects is directly correlated with an animal's ability to successfully perform a task [16, 18]. Finally, the extent to which acoustic features of the foreground and background are enhanced and suppressed, respectively, is governed by the STRF parameter update equations (box C). This may have parallels with neurotransmitters from NB and/or VTA and how they shape STRF sensitivity to specific target frequencies or spectro-temporal modulations.

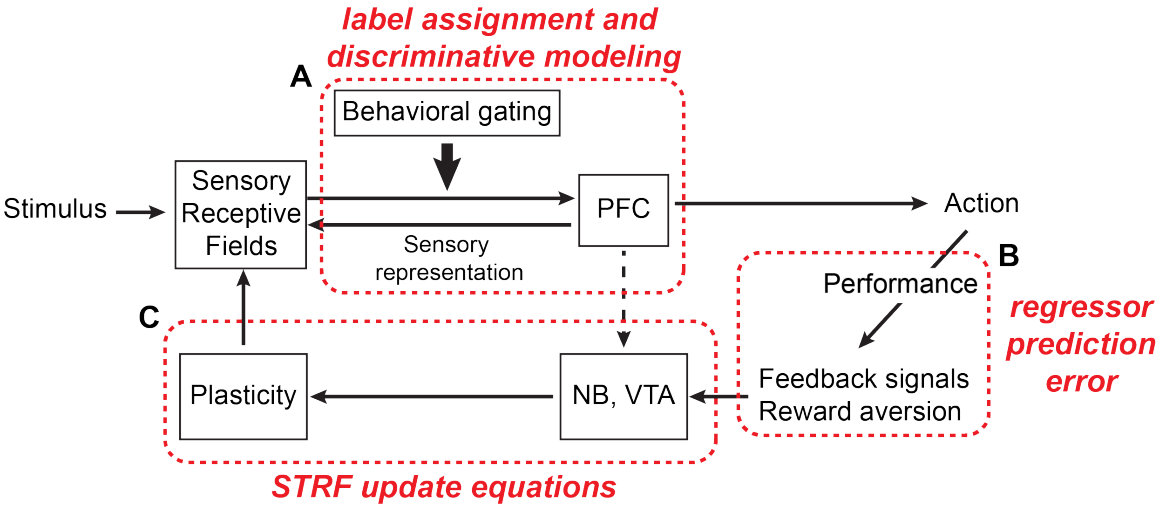

Figure 3.8: Simplified schematic of anatomical circuits thought to be involved in attention-driven auditory cortical plasticity (adapted from [144]). Refer to text for details.

3.4.2 Relationship between the Feature- and Object-Based Models

Modulo use of a spectro-temporal mask and choice of neural firing rate model (i.e., 1D vs. 2D convolution), the receptive field adaptation mechanisms predicted by the Feature- and Object-Based Models are at their core comparable. This is clear by directly comparing the STRF update

equations given in Eqs. 3.6 and 3.10, where the primary difference between the two is the use of stimulus phase during adaptation. More generally, to distinguish object- from feature-based attention, we considered separately the spectro-temporal magnitude and phase profiles of the observed acoustics. This allowed us to quantify the notion that object-based attention requires that cognitive resources be directed towards an *abstracted* representation of task-relevant sound classes, represented here by the collection of modulations that comprise the acoustic foreground. However, object-based attention is certainly not restricted to act merely on the Fourier domain representation of sound, since attention can act at even higher levels of abstraction, for example, by being directed to an individual melody in an orchestra, the prosodics of a target speaker at a cocktail party, or a bird watcher listening for a specific species call in nature. Furthermore, while we have drawn a clear distinction between the notions of feature- and object-based attention, the existence of such a clear difference between the two is still the subject of debate (see, e.g., [137] and [54]). Nevertheless, the proposed framework provides a means to evaluate both hypotheses as more physiological and behavioral results become available.

For the Object-Based Model, the choice of 2D convolution for modeling neural firing rate was motivated by prior work that suggests that such a representation is sufficient to capture a variety of aspects of sound perception such as speech intelligibility and timbre representation [3, 146, 147]. Of course, it may be possible to adapt the STRF modulation profiles using a 1D firing rate model. However, we feel that our 2D formulation is simpler, intuitive, and, more importantly, reflects the fundamental mechanism implied by neurophysiological studies, namely, that STRFs reorient themselves to act similar to a contrast matched filter in the Fourier domain for complex spectro-temporal tasks [50, 77].

Under what circumstances does a listener employ the Feature- and Object-based models? We hypothesize that this decision depends on task, and that the final choice is made empirically based on the behavioral outcomes of either strategy. Again, the key distinction between the two models is the use of stimulus phase in the STRF update equations (Eqs. 3.6 and 3.10). So for

tasks where exploiting differences in phase is important, like tone discrimination (spectral phase) or speech recognition (temporal phase), the Feature-Based Model will be employed. Conversely, for tasks where the task-relevant classes are distinguished largely based on differences in spectro-temporal modulation profiles, as with conspecific vocalizations versus ambient environmental noise, the Object-Based Model will be employed. Of course, it is also possible that the predictions of both models, coupled with other sources of contextual information, are combined to make an overall decision. Future work should explore how exactly to combine the models, as well as how to quantify and incorporate context into the current framework.

3.4.3 Related work

Our framework was conceived in the spirit of the approaches of [80] and [76], where they proposed discriminative cost functions that quantified the computational goal of task-related plasticity in the auditory system. However, as discussed earlier, these models lacked two important components: (1) a guarantee that optimal solutions capture task-valence (i.e., when the behavioral meaning of target and reference are flipped, the direction of plasticity is also flipped) and (2) the ability to adapt STRFs based on an abstracted representation of the stimulus. Our framework directly addresses these issues, predicting a qualitative correspondence with existing physiological data, and addressing stimulus phase-invariant adaptation of STRFs via their modulation profiles—which, interestingly, has also been observed in visual cortical area V4 [50].

More generally, however, a strong connection exists between our approach and a recently proposed framework of top-down attention in vision. [70] describe an optimal attention framework that accounts for a variety of attention-driven plasticity effects in visual cortex for discrimination and search tasks, and yields predictions that qualitatively explain a broad range of attentional mechanisms that depend on task type and difficulty. Their framework is based on deriving a set of filter gain and tuning parameters that optimize a task-dependent objective function (e.g., discrimination or visual search) and prescribing an appropriate optimization procedure. The Borji and Itti vision

model shares a lot of parallels with the proposed scheme that was mainly inspired by task-driven effects observed in mammalian A1 and tailored to the specific particularities of the auditory system. The strong concordance between the two frameworks supports the notion that the discriminative cost function proposed here for task-driven plasticity applies broadly across many sensory modalities.

3.4.4 Other model considerations

For both models presented here, we selected the hyperparameters (C, λ) based on what we felt yielded a reasonable correspondence with published or expected physiological results. However, just as average plasticity patterns observed in animal studies vary based on factors like motivation, level of alertness, and satiation, the plasticity patterns predicted by the model vary with choice of (C, λ) (see Fig. B.1 in the appendices). The specific values of these coefficients are not critical (since they depend factors like the amount of stimulus used and normalization applied to the STRFs and stimuli), but their importance from a modeling perspective is that they provide a mechanism for trading off between the neurophysiologically implied coding heuristics of discriminability and stability. Specific values of these parameters could be determined using cross-validation on available behavioral results and measured passive/active STRFs, but this is beyond the scope of this study.

We have interpreted the notion of a contrast filter as referring to adaptation of primary cortical receptive fields that selectively enhance and suppress acoustic features of the foreground and background, respectively [75, 148]. This is captured in our model by the addition of nonnegativity constraints on the regressor coefficients. In the development of the Object-Based Model, we noted that the spectro-temporal mask—which guaranteed local plasticity in the Feature-Based Model—was omitted for technical reasons (see the supplementary text in Appendix B). In short, including a mask in the firing rate model introduces a sign ambiguity in the gradient w.r.t. the modulation profiles and as a result, even with nonnegativity constraints on the regressor, we are no longer guaranteed that target modulations will be enhanced and reference modulations be suppressed. Thus, plasticity predicted by this version of the model is not strictly consistent with our interpretation of

the contrast filtering hypothesis. However, data from [52] suggest that while on average target (reference) responses are enhanced (suppressed), there are many instances at synapses from A1 through prefrontal areas where opposite patterns are observed (i.e., target responses suppressed and vice versa for reference responses). This may reflect similarly reversed underlying receptive field plasticity patterns. Thus, just because the model enforces constraints that guarantee strict consistency with the contrast filtering hypothesis, versions of this model without such constraints will still yield interpretable results, with a modular model structure that can be mapped to circuits likely involved in attention as described earlier in the Discussion.

Chapter 4

An Adaptive Framework for Speech Activity Detection Based on Top-Down Auditory Attention

4.1 Introduction

As discussed in Chapter 3, part of the reason the auditory system performs well in tasks like tracking a voice in a noisy cocktail party or otherwise parsing complex acoustic scenes is that the internal representation is not static, but can dynamically adapt using top-down attentional cues. In particular, using our proposed framework, we showed how the tuning properties of the STRFs can vary to improve discrimination between foreground and background sounds, subject to biologically plausible constraints. Because these changes enhanced representation of foreground sounds while actively suppressing the response to the background, this makes the framework especially attractive for application to automated sound processing systems that handle noisy or highly confusable signals. In this chapter, we apply the Object-Based formulation of our proposed attentional framework to

the problem of extracting information from speech signals corrupted by additive noise.

Of particular importance for human communication is the ability to selectively attend to and track speech, a task at which listeners are especially adept in noisy environments [149]. Attention-driven adaptation strategies that improve the separation between foreground speech and background nonspeech sounds are thus particularly attractive for automated speech processing tasks. Speech signals can be richly characterized by their spectro-temporal modulation content, and so the Fourier domain is a natural space for developing such strategies. This is because a number of speech features can be expressed jointly in the spectro-temporal modulation domain, including voicing state, phoneme identity, formant trajectories, and syllable rate [3,102,146,150]. Furthermore, sounds that have considerable overlap in time-frequency may in fact be disjoint in the modulation domain, leading to methods for signal denoising and enhancement [151]. Finally, as we argued in the previous chapter, modulation-domain adaptation may reflect a general form of object-based auditory attention. This is because adapting the Fourier magnitude profile, which characterizes the strength of spectro-temporal modulations present in the signal, separately from its phase profile, which characterizes the relative timing of these modulations, is akin to processing an *abstracted* representation of the signal, an important component of object-based attention [6, 7, 54, 152].

In this chapter we explore application of the Object-Based Model of attention to the challenge of speech activity detection in noisy environments. We begin by describing some basic adaptation results using a speech target in the context of a noisy nonspeech reference, and show how the STRFs reorient themselves to enhance the spectro-temporal modulations of speech while suppressing those associated with a variety of nonspeech sounds. We next explore application of features derived from the adapted STRFs in a speech activity detection (SAD) task. The Object-Based model is a natural fit for SAD, and we demonstrate improved detection performance in mismatched noise scenarios with respect to the unadapted ensemble and a recently proposed baseline. Finally, to better understand how STRF adaptation affects the representational quality of attended speech, we consider a stimulus reconstruction task similar to those recently considered in a variety of neuro-

physiological studies. We show that stimuli reconstructed from STRFs adapted using the proposed framework yield a higher fidelity representation for speech in clean and additive noise conditions using a variety of objective and perceptual measures. Overall, the results suggest that the Object-Based Model formulated in the modulation domain can yield a high-fidelity, noise-robust representation of the attended source that is applicable to automated speech processing tasks.

4.2 STRF Plasticity for Noisy Speech Detection

We first applied the Object-Based model (See Chap. 3.3.5 to simulate a scenario where a listener directs their attention to speech in additive noise environments. We used clean speech from the TIMIT corpus [82] and equal amounts of `white`, `babble`, `street`, `restaurant`, and `f16` noise from the NOISEX-92 corpus [108]. The target class was constructed to contain equal amounts of clean and noisy speech (at 5 dB SNR) whereas the reference class contains pure noise. The use of clean and noisy speech in the target class reflects the notion that a listener has prior knowledge of speech in both clean and moderately noisy environments. We used audio samples approximately 3 seconds in length, applied pre-emphasis, and standardized each waveform to be zero-mean and unit variance, and we used approximately 5 minutes of audio for each class. Next, for each audio sample we computed an auditory spectrogram, applied cube-root compression, and downsampled along the frequency axis to 32 channels. Finally, the 2D discrete Fourier Transform (DFT) was applied to 250 ms segments of spectrogram, followed by the modulus operation to obtain input tokens $|S_m(\omega, \Omega)|$ for the adaptation algorithm. The tokens were scaled to have unit variance for each class as this seemed to improve convergence time of the adaptation algorithm.

We next selected an initial ensemble of $K = 50$ STRFs uniformly at random from the large physiological ensemble described in Chap. 3.3.1. The use of a larger ensemble was computationally challenging because the number of parameters involved in the optimization was prohibitively large. Furthermore, we found that random samplings 50 STRFs were sufficient to tile the relevant mod-

ulation space of speech tokens given the redundant and overcomplete nature of neurophysiological receptive fields [153, 154]. Next, each STRF was interpolated along the time and frequency axis to match the temporal and spectral sampling of the input tokens (i.e., 100 Hz temporal and 6.04 cyc/oct spectral, respectively). We also assumed that each STRF had a base frequency of 90 Hz spanning 5.3 octaves. We scaled each STRF to have unit Euclidean norm, and applied the 2D DFT followed by the modulus operation to obtain the initial set of modulation profiles $\mathcal{H}_0 := \{|H_k^0(\omega,\Omega)|\}_{k=1}^K$. We assume that this set represents a "passive" listening state, one where attention is not engaged. Note that for adaptation we only need to consider the first two quadrants of the DFT since for real-valued input $|H_k^0(\omega,\Omega)| = |H_k^0(-\omega,-\Omega)|$. Finally, for visualizing the adapted STRFs in the original time-frequency domain, we use the phase of the original passive filters.

Upon optimizing the cost-function of the Object-Based Model (see Chap. 3.3.5), we obtain a set of attention-modulated modulation profiles $\mathcal{H}_A$. To characterize the effect of attention on the neural ensemble, we consider the average difference between the adapted and initial modulation profiles by computing

$$\Delta MTF(\omega,\Omega)\rangle := \langle |H_k^A(\omega,\Omega)| - |H_k^0(\omega,\Omega)| \rangle_k$$

In this way, one can visualize which modulations of the speech and nonspeech stimuli are enhanced suppressed, respectively.

The results of this analysis are shown in Fig. 4.1A, and illustrate that the overall effect of attention is to increase population sensitivity to slower modulations, with the effect being stronger for downward vs. upward moving modulations (i.e., the right- vs. the left-half planes, respectively), while suppressing sensitivity to faster modulations away from the origin. This pattern was also found for other random selections of the initial ensemble $\mathcal{H}_0$. The magnitude of the effect depends on choice of model hyperparameters, becoming stronger for decreasing λ and increasing C (data not shown). Finally, shown next in panels B and C are the adaptation patterns of two individual neurons, illustrating how the neurons broaden and reorient themselves to better focus on upward

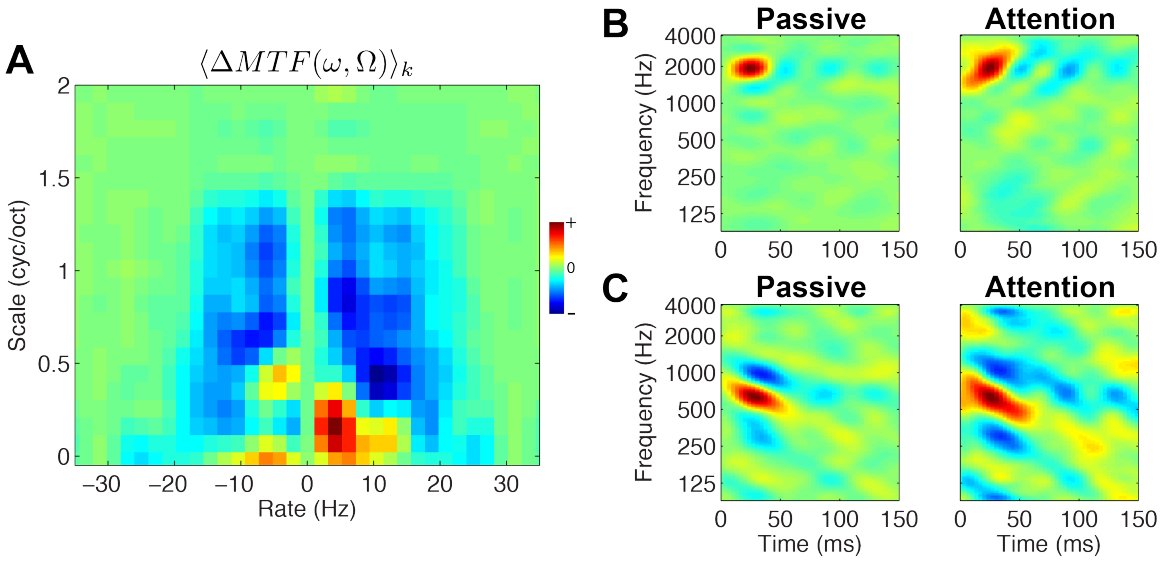

Figure 4.1: Effect of the object-based attentional model for speech-in-noise detection task. (A) Population effects were quantified by comping the average difference modulation profile $\Delta MTF(\omega, \Omega)$ and show that the object-based model tends to increase sensitivity to slower modulations close to the origin while suppressing faster modulations away from the origin. (B and C) Individual examples illustrating how attention induces the STRFs to reorient themselves according to the task. Results shown for $\lambda = 10^{-1}, C = 10^{2.25}$.

and downward modulations as suggested by panel A. In the next section, we explore the extent to which these adaptations improve performance on a speech activity detection task.

4.3 Top-down Attention for Speech Activity Detection

Speech activity detection refers to the task of assigning a speech or nonspeech label to each sample in an observed audio recording and is a fundamental first step in a number of automated sound processing applications. For example, in automatic speech recognition tasks, one should transcribe only speech events in the observed audio, respectively. In low-noise environments with a close-talking microphone, SAD can usually be solved using traditional measures like signal energy, zero-crossing rate, and pitch (see, e.g., [155–157]), but performance rapidly degrades in noisy, reverberant, and free-field microphone conditions.

However, research in the past decade has begun to focus on the issue of noise-robustness

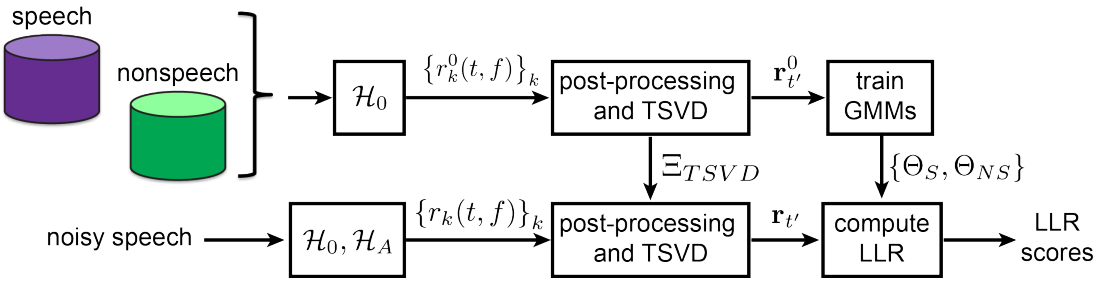

Figure 4.2: Overview of the proposed SAD system.

SAD, with successful approaches relying on some prior knowledge about the statistics of acoustic features from speech versus nonspeech sounds. State-of-the-art results are typically obtained using data driven methods based on Gaussian Mixture Models (GMMs) and deep convolutional neural networks [158, 159], however such systems are typically trained on hundreds of hours of audio.

We take a different approach to SAD, and consider to what extent attention-driven "retuning" of sensory filters can improve SAD performance. We did not optimize choice of the initial filter set in order to maximize SAD performance (see Discussion). Instead, here we focus on simulating a scenario where a listener has prior knowledge about the statistics of clean and noisy speech, as quantified by a GMM, and adapts their feature representation in noisy environments to attend to a target speech source. In particular, we hypothesize that an ensemble of STRFs adapted according to the proposed attentional framework will improve performance in a SAD task. Here we will show that a system trained using features derived from a passive STRF ensemble is substantially improved when using attention-modulated features.

4.3.1 Experimental setup

An overview of the proposed SAD system is shown in Fig. 4.2. For training (top row), we use clean speech and a variety of nonspeech samples to extract a set of features from the passive ensemble $\mathcal{H}_0$, yielding ensemble responses $\left\{r_k^0(t,f)\right\}_k$. The firing rates are computed as in Eq. 3.7, with 128-channel auditory spectrograms and cube-root compression applied. The STRFs $h_k(t,f)$ were

also interpolated to span the full 128 channels. Computation of the ensemble response is followed by a series of post-processing and dimensionality reduction steps. First, the responses are full-wave rectified and averaged over one-second intervals every 50 ms, yielding a three-dimensional tokens $\mathbf{R}_0(t', f, k)$. We next apply dimensionality reduction using the tensor singular value decomposition (TSVD) [160], projecting the tokens to a subspace that retains 99.9% of the variance along each dimension, and stack the reduced-dimension tokens to obtain column vectors $\mathbf{r}^0_{t'}$. We then standardize each vector to have zero-mean and unit-variance. Finally, we fit Gaussian mixture models (GMMs) to the observed speech and nonspeech tokens, yielding model parameters (i.e., weights, means, and covariance matrices) Θ_S and Θ_{NS}, respectively.

For testing (bottom row), features are extracted from observed noisy speech utterances using both passive and attention-modulated STRFs $\mathcal{H}_0$ and $\mathcal{H}_A$, respectively. We apply post-processing and dimensionality reduction as described above, and compute the log-likelihood ratio (LLR) of speech versus nonspeech using the GMMs trained in the passive conditions. We evaluate system performance by sweeping a threshold on the LLR scores, labeling tokens that exceed the threshold as speech, and those below as nonspeech. Using ground truth labels, for a given threshold value we compute miss and false alarm probabilities, p_M and p_{FA}, respectively, and compile these error probabilities across all thresholds to yield a detection error tradeoff (DET) curve [161]; an example DET curve is shown ahead in Fig. 4.3A. To summarize performance of the system, we compute the equal-error rate (EER), i.e., the threshold setting that yields $p_M = p_{FA}$. A system that performs well has small p_M and p_{FA} across a broad range of thresholds, hence (1) the corresponding DET curve will be close to the origin and (2) EER will be small.

4.3.2 Results

To test the hypothesis that attention-modulated STRFs can improve detection of speech in unseen noisy environments, we built a GMM-based SAD system and compare the performance of between features derived from the simulated passive and active attentional states. For the passive and

attention-modulated STRFs, we consider three random draws from the large physiological STRF set as well as a range[1] of model hyperparameters C; we report results for the STRF ensemble that yields the best performance. We trained our GMM SAD system using clean speech from the TIMIT corpus and nonspeech samples from the BBC Sound Effects Library [83]. The nonspeech set comprises an equal amount of audio from a range of acoustic classes[2]. For both the speech and nonspeech categories, we use 7500 one-second tokens, or approximately 2.1 hrs of audio, and reduce the dimension of extracted features via TSVD from (128 frequency channels $\times$ 50 STRFs) to 40-dimensional column vectors. Finally, we fit 32-mixture GMMs to the speech and nonspeech categories.

We evaluated our system in unseen noise cases using audio from the QUT-NOISE-TIMIT corpus, a database specifically designed for evaluating different SAD algorithms [162]. The corpus is formed by mixing clean TIMIT utterances with realistic background noise covering four noise scenarios[3] (`STREET`, `CAFE`, `CAR`, and `HOME`) with two unique locations per noise type at various SNRs. For our experiments, we selected ten utterances (each 60-seconds in length) at random from each noise condition and SNR, ensuring that there was no overlap between TIMIT utterances seen in training and those used in testing. For a baseline comparison, we use the statistical model-based likelihood ratio test of Tan *et al.* that leverages harmonicity cues to improve detection of speech in noisy environments [163]; this approach has been shown to work well in a variety of noise conditions.

Shown in Fig. 4.3A is an example DET curve for the `STREET` noise for the baseline and proposed system using passive and attention-modulated STRFs. The DET curve is obtained by pooling LLR scores across SNRs $\in \{-5, 0, +5, +10, +15\}$ dB. Because the DET curve for the attention-modulated ensemble $\mathcal{H}_A$ is closest to the origin with no overlap with the other curves, it represents clear improvement over the baseline and the passive ensemble $\mathcal{H}_0$ across all SNRs, with an absolute reductions in EER of approx. 11% and 3%, respectively.

[1] We found it was sufficient to fix λ and explore a range of values for C.

[2] For this study we used the `Emergency`, `Foley`, `Industry and Machines`, `Technology`, `Transportation`, and `Water` classes.

[3] There is also a fifth `REVERB` condition, but for our experiments it is ignored.

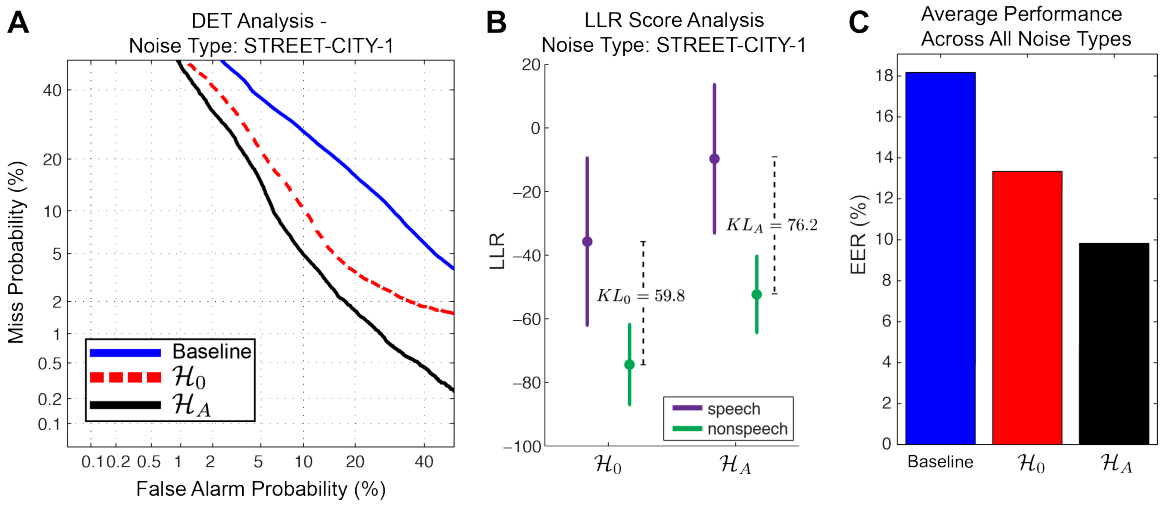

Figure 4.3: SAD Results. (A) DET curves for the `street` noise condition, computed by pooling scores across all SNRs, for the baseline, passive, and attention-modulated STRFs. (B) Visualization of LLR score distributions (as mean and standard deviation) for the passive and attention-modulated ensembles, showing how use of $\mathcal{H}_A$ improves separation between the speech and nonspeech LLR scores. (C) Average EER over all noise conditions. Attention-modulated STRF ensembles are reported for $\lambda = 10^{-1}, C = 10^{2.0}$.

To better understand how the attention-modulated STRFs improve performance, Panel B shows an analysis of the distribution of LLR scores for the STRF ensembles with respect to speech and nonspeech categories. These results show that under the GMMs trained on the passive STRF features, use of the attention-modulated STRF ensemble increases the overall likelihoods of speech and nonspeech. However, despite this added bias to the scores, there is an overall improved separation between the speech and nonspeech distributions as computed using the Kullback-Leibler divergence ($KL_0 = 59.8$ and $KL_A = 76.2$ assuming Gaussian-distributed scores). Similar improvements are found across the other noise scenarios.

Finally, panel C summarizes the overall performance of the various SAD configurations in terms of average EER across all noise conditions, showing that the attention-modulated STRFs improve over the baseline and passive STRF results by an absolute 8% and 3.5%, respectively.

4.4 Further Analysis

The previous section showed empirically that the proposed STRF adaptation framework improved detection of speech in unseen noisy conditions by increasing the separability between the LLR scores of speech and nonspeech (Fig. 4.3B). However, we sought to better understand how the adapted STRF ensemble improved the representational fidelity of attended speech. One way to assess the ability of a neural ensemble to encode features of the attended source is to use a stimulus reconstruction approach. By reconstructing the observed input to the ensemble, one can assess how the features of the input are encoded by the population. One can then vary the attentional state of the ensemble (i.e., passive vs. attentive) and compare the reconstructions with the original stimulus.

The stimulus reconstruction approach has been successful in a number of neurophysiological studies. The approach was pioneered in studies of the fly visual system [164, 165] and has been used to study feature encoding in visual [166, 167] and auditory cortical circuits [168, 169]. Of particular interest are recent studies that have shed light on how cortex represents imagined speech [170] and the nature of how top-down auditory attention influences representation in cortical circuits [171–173].

In this section, we explore the stimulus reconstruction approach as it relates to the challenge of speech-in-noise detection, using the approach outlined by Mesgarani et al. [168]. We hypothesize that adapting an STRF ensemble according to the proposed model yields a higher fidelity representation of the attended stimulus, and we explore this via reconstruction experiments in clean and additive noise conditions using objective and perceptual measures.

4.4.1 Stimulus reconstruction

In physiological studies, neural firing rate is typically modeled as

$$\tilde{r}_k(t) = \sum_{f=1}^{F} h_k(t,f) *_t s(t,f)$$

where $h_k(t, f)$ is an STRF, $s(t, f)$ is the stimulus, F is the number of frequency channels, $t \in [1, T]$, and $*_t$ denotes convolution in time. To reconstruct an input stimulus from observed neural firing rates we use the linear form

$$\hat{s}(t, f) = \sum_{k=1}^{K} \sum_{\tau=1}^{\tau_M} g(k, \tau, f) \tilde{r}_k(t - \tau) \tag{4.1}$$

where $\tau_M > 0$ is the (user-defined) temporal extent of the reconstruction filters and $\{g(k, \tau, f)\}$ is a collection of inverse mapping functions [168].

Eq. 4.1 implies that individual frequency channels are independent of one another and hence we can compactly write

$$\begin{aligned} \hat{s}_f(t) &= \sum_{k=1}^{K} \sum_{\tau=1}^{\tau_M} g_f(\tau; k) \tilde{r}_k(t - \tau) \\ &= \mathbf{g}_f^T R \end{aligned}$$

where

$$\mathbf{g}_f = \text{vec} \left[\begin{pmatrix} g_f(1,1) & g_f(1,2) & \cdots & g_f(1,K) \\ g_f(2,1) & g_f(2,2) & \cdots & g_f(2,K) \\ \vdots & \vdots & & \vdots \\ g_f(\tau_M,1) & g_f(\tau_m,1) & \cdots & g_f(\tau_M,K) \end{pmatrix} \right]$$

and

$$R = \begin{pmatrix} \tilde{r}_1(1) & \tilde{r}_1(2) & \cdots & \tilde{r}_1(\tau_M) & \cdots & \tilde{r}_1(T) \\ 0 & \tilde{r}_1(1) & \cdots & \tilde{r}_1(\tau_M - 1) & \cdots & \tilde{r}_1(T-1) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \tilde{r}_t(1) & \cdots & \tilde{r}_1(T - \tau_M) \\ \tilde{r}_2(1) & \tilde{r}_2(2) & \cdots & \tilde{r}_2(\tau_M) & \cdots & \tilde{r}_2(T) \\ 0 & \tilde{r}_2(1) & \cdots & \tilde{r}_2(\tau_M - 1) & \cdots & \tilde{r}_2(T-1) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \tilde{r}_2(1) & \cdots & \tilde{r}_2(T - \tau_M) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \tilde{r}_K(1) & \cdots & \tilde{r}_K(T - \tau_M) \end{pmatrix}$$

The vec$(\cdot)$ operator performs a column-wise stacking of the input matrix. Furthermore, defining the inverse mapping matrix $G := [\mathbf{g}_1 \cdots \mathbf{g}_F]$, we can write $\hat{S} := \hat{s}(t, f) = G^T R$.

One way to arrive at an optimal reconstruction $\hat{S}$ is to determine the matrix G^* that solves the least-squares problem

$$G^* = \arg\min_G ||S - \hat{S}||_F^2$$

where $S := s(t, f)$ is the observed stimulus. This closed-form solution is readily obtained as

$$G^* = C_{RR}^{-1} C_{RS} \tag{4.2}$$

where $C_{RR} = RR^T$ is the response autocorrelation matrix and $C_{RS} = RS^T$ is the stimulus-response correlation matrix. Because of correlations in the stimulus and the fact that there are some redundancies in terms of filter shape the neural ensemble, the observed firing rates consequently yield redundancies in the rows of R. Thus, it is often the case that C_{RR} is poorly conditioned, necessitating the use of some form of regularization to properly invert the response autocorrelation matrix. Here we use the subspace regression approach proposed by Theunissen et al. [14]. Since

C_{RR} is real-symmetric, it can be expressed as $C_{RR} = U\Sigma U^T$ where U is a matrix of eigenvectors, $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_{K\cdot\tau_M})$, and σ_i are the eigenvalues of C_{RR}. Because of the redundancies in the rows of R, it generally holds that $\text{rank}(C_{RR}) < K \cdot \tau_M$ and thus we should ignore the eigenvectors corresponding to small eigenvalues (otherwise they tend to introduce noise once C_{RR} is inverted). We set eigenvalues smaller than a pre-defined threshold η to zero.

4.4.2 Experimental setup

We consider two measures to evaluate the quality of a given reconstruction from an inverse mapping obtained by Eq. 4.2. The first is the temporally averaged mean-square error between the original and reconstructed spectrogram, defined as $MSE := \langle ||s(t,f) - \hat{s}(t,f)||_F^2 \rangle_t$ and serves as an objective measure of reconstruction quality. The second is a perceptual comparison between the original time-domain waveform and and synthesized version obtained using the Gaussian convex projection algorithm [2,3]. The comparison between the waveforms is made using the ITU standard Perceptual Evaluation of Speech Quality (PESQ) measure [174]. PESQ ranges between 1 and 5 and correlates well with listener-reported mean opinion scores of perceptual quality, with higher scores indicating higher quality.

To study how reconstruction performance varies as a function of attentional state, we use the passive and attention-modulated STRF ensembles $\mathcal{H}_0$ and $\mathcal{H}_A$ to obtain optimal inverse mappings G_0 and G_A, respectively. We use 350 clean speech utterances from the TIMIT `train` corpus (approx. 17.5 minutes) to learn the inverse mapping matrices. The neural responses $\tilde{r}_k(t)$ are also standardized to have zero-mean and unit variance prior to obtaining G_0 and G_A, and these parameters are applied to subsequent reconstructions. For a given inverse mapping, we also consider a range of inverse filter lengths spanning $\tau_M \in [50, 750]$ ms and eigenvalue thresholds $\log_{10}\eta \in \{-9, -6, -3\}$. Results are reported here for ensembles that achieve minimum average mean-square reconstruction error on a test set of 100 clean speech utterances from the TIMIT `test` corpus. For synthesizing time-domain waveforms, we first apply a $\max(\cdot, 0)$ nonlinearity to

synthesized spectrograms, followed by a maximum of 30 iterations in the Gaussian convex projection algorithm.

4.4.3 Results

Shown in Fig. 4.4A are examples of reconstructions from clean utterances obtained using the passive and attention-modulated STRF ensembles. We first find that both reconstructions are somewhat noisy, with G_0 yielding a distinct temporal distortion whereas G_A introducing spurious patches of spectro-temporal energy. However, both reconstructions are sufficient to capture the broad prosodic characteristics of the reference spectrogram, with good qualitative matches between pitch variations, syllabic rate, and broad formant dynamics over time. Furthermore, it is clear that G_A yields a reconstruction with better spectral resolution, since the harmonic peaks during sections of voicing are far more pronounced as compared with G_0. Next, Panel B shows that across the test set, the attention-modulated ensemble yields an objectively better reconstruction, with G_A yielding a significantly lower reconstruction error as compared to G_0 (t-test, $p \approx 0$). Finally, the perceptual analysis in Panel C shows G_A yields a significantly higher quality waveform synthesis compared G_0 (t-test, $p \approx 0$). Of course, while PESQ values between 2-3 are generally considered somewhat noisy, informal listening tests confirm that the synthesized waveforms are nevertheless intelligible, with G_A conveying a better percept of voicing and pitch.

We also explore the extent to which the passive and attention-modulated STRF ensembles encode information about the attended source in additive noise conditions. Here we consider a test set of 100 utterances from the TIMIT *test* corpus corrupted by additive noise from the NOISEX-92 corpus at a variety of SNRs. In addition to the `babble` and `street` noises used in training, we also consider the unseen noise classes `airport`, `buccaneer1` (a type of fighter jet), `factory1`, and `m109` (a type of tank). For each noisy utterance we reconstruct the spectrogram using the inverse mappings G_0 and G_A. We then quantify reconstruction quality using the time-averaged mean-squared error between the clean reference spectrogram and the noisy reconstruction. These results, averaged across

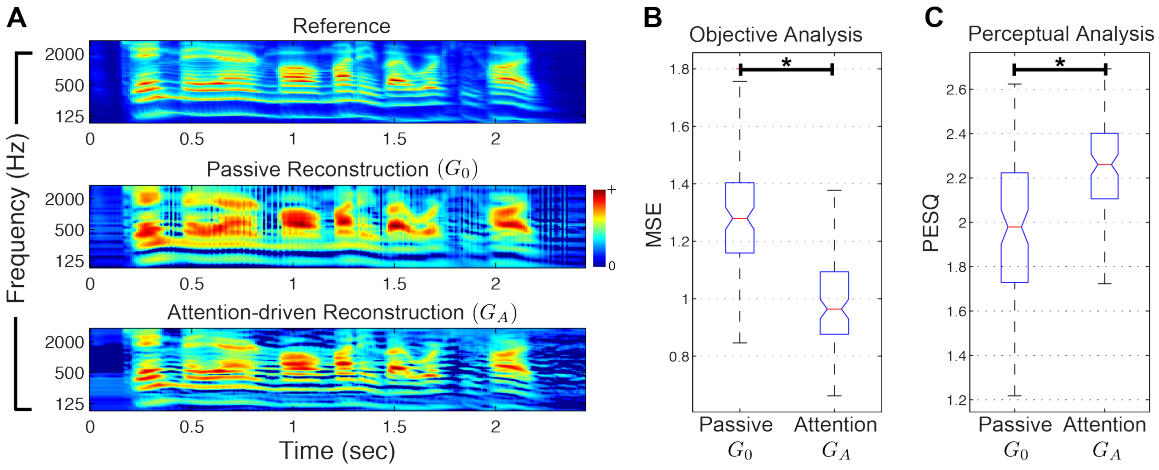

Figure 4.4: Analysis of clean speech reconstructions. (A) A reference spectrogram (top) is compared with reconstructions obtained from the passive (middle) and attention-modulated STRF ensembles (bottom). (B) Objective analysis shows that G_A yields a significantly better reconstruction compared to G_0. (C) Perceptual analysis shows that G_A yields a significantly higher quality waveform synthesis compared to G_0 ($*$: t-test, $p \approx 0$). Results shown for $\lambda = 10^{-1}, C = 10^{2.0}$.

all noise types, are shown in Fig. 4.5. It is clear that while reconstruction quality degrades with increasing noise, in all SNR cases the attention-modulated ensembles yield, on average, a higher quality reconstruction with respect to the clean references. We found no differences between average MSE for the seen vs. unseen noise cases.

In summary, the results of this section suggest that the proposed model of attention-driven adaptation induces STRF changes that a facilitate higher-fidelity representation of attended speech in clean and noisy environments. This lends further insight as to how such an adaptation strategy is able to improve detection of speech in noisy environments.

4.5 Discussion

In this chapter, we applied a model of top-down attention that acts in the modulation domain to simulate the scenario of a listener "directing their attention" to better track speech in noisy acoustic environments. We first described how an ensemble of passive STRFs would vary in the case of speech vs. nonspeech. We showed that the model induces STRF plasticity that strengthens

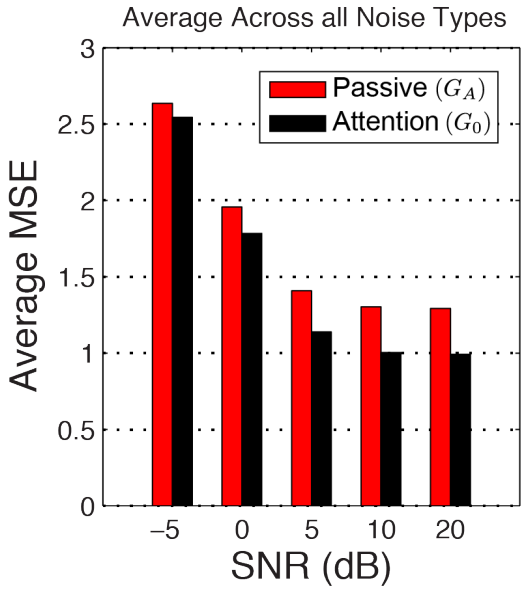

Figure 4.5: Analysis of noisy speech reconstructions. Reconstruction quality degrades with increasing noise. However, in all SNR cases, the attention-modulated ensembles yield, on average, a higher quality reconstruction with respect to the clean references. Results shown for $\lambda = 10^{-1}, C = 10^{2.0}$.

relatively slow spectro-temporal modulations close to the origin (in the rate-scale domain) while simultaneously suppressing faster modulations away from the origin. We then showed how use of the adapted STRFs improves the separation between the representation of speech and nonspeech sounds, resulting in a substantial performance gain in a speech activity detection task across a variety of previously unseen noise types. Finally, we explored, via stimulus reconstruction experiments, the extent to which the passive and attention-modulated STRF ensembles captured the salient features of the target speech source. These results showed how the use of attention can improve the representation of a speech target in clean and noisy conditions, as confirmed by objective and perceptual measures. This helped to shed light as to exactly how the representation improved to help facilitate the detection of speech in noise. The overall results suggest that STRFs adapted according to a biologically motivated model of top-down attention can form a noise-robust representation of sound that is applicable to automated speech processing applications.

The Object-Based model predicts that attending to speech versus nonspeech distractors

enhances sensitivity of the STRFs to "slower" spectro-temporal modulations. This is illustrated by the average difference modulation profile in Fig. 4.1. Increased slowness in the modulation domain is realized in the time-frequency domain as an overall broadening and reorientation of the STRFs, and reflects an enhancement of the modulations known to characterize speech and other natural sounds [3, 14, 102, 146]. The fact that we obtain improved SAD results using "slower" filters is consistent with other strategies that concentrate the feature extraction pipeline to the range of speech-specific modulations [103,150,175,176]. Moreover, the STRF adaptation patterns observed here are broadly compatible with traditional signal processing schemes that emphasize slow modulations for improving noise robustness in speech tasks [177], and our framework is broadly consistent with data-driven approaches for feature extraction for speech applications [178]. The distinction here is that our approach adapts the filter shapes "on the fly" and, as our SAD results suggest, such changes can be compatible with an existing statistical model to improve task performance.

As stated earlier, attending to speech in noisy environments is a critical component of human communication, and is a task at which listeners are especially adept. Based on this, and the fact that reliably detecting speech-containing regions is a critical first step in many common speech applications, the task of speech activity detection is a natural fit for our framework. Our goal in this study was not to build a state-of-the-art SAD system *per se*, but to instead focus on the design and understanding of the impact of an adaptive spectro-temporal algorithm for speech that is grounded in auditory neurophysiology. However, we expect that performance of the framework would improve by optimizing choice of the initial filter set, and we leave this fine-tuning of the system for future work. That being said, many robust approaches to SAD have been proposed over the years that perform well in noisy and reverberant environments (see, e.g., [155, 163, 179–181]). In particular, when large amounts of training data are available, extraordinary results can be achieved for the SAD task on difficult corpora using high-order GMMs [158] and convolutional neural networks [159, 182]. Nevertheless, these systems continue to be challenged by the challenges of nonstationary interference, mismatch between expected and observed acoustics, and limited training data, and it is our hope

that the modulation-domain adaptation framework presented here can be leveraged to improve these approaches.

More generally, the increasing availability of mobile sound processing applications has resulted in a significant increase in the variety of acoustic environments, communication channels, and noise conditions encountered by existing systems. Consequently, this necessitates signal processing strategies that must gracefully accommodate these factors to maintain state-of-the-art performance. We contend that because nature has converged to a robust solution for handling unseen and noisy acoustics, there is much to leverage from the auditory neurophysiology when designing automated sound processing systems. Generally speaking, cortically inspired feature representations based on spectro-temporal receptive fields underlie a number of successful approaches to noise robust speech activity detection [150], speech and speaker recognition [103, 175, 176, 183], and auditory scene classification [184]. The present study, in concert with other recent work in our lab [152, 185, 186], represents an extension of this methodology by incorporating the cognitive effects of dynamic, task-driven sensory adaptation as part of the feature extraction pipeline. It is our belief that new and existing systems can only benefit by incorporating the adaptive mechanisms as outlined in this chapter.

Chapter 5

Conclusions

5.1 Overall conclusions

This thesis has explored the computational objectives underlying the formation of auditory representations and top-down, attention-driven adaptation in the central auditory system. The first part of the thesis explored the consequences of a sustained firing criterion on the shapes of STRFs, demonstrating the emergence of a highly-structured receptive fields that capture the statistics of natural sounds. We quantified the individual- and population-level characteristics of these STRFs using a variety of measures, and made a detailed comparison between the model predictions and known neurophysiological results. We also explored how the sustained firing code compared to other popular strategies based on sparsity and slow feature analysis. Finally, we showed how the emergent ensemble could be applied to achieve noise-robustness in an ASR task, outperforming a state-of-the-art baseline.

In the next part of the thesis, we proposed a discriminative cost function whose optimal solution explains how STRFs adapt to enhance and suppress features of the acoustic foreground and background, respectively. Importantly, the adaptation patterns predicted by the framework are consistent with the contrast filtering hypothesis thought to act in A1, and our predictions have a

close correspondence with known neurophysiological data. We next showed that a generalization of the framework, which acts on the spectro-temporal dynamics of task-relevant stimuli, induced plausible predictions for tasks that have yet to be experimentally measured. We also argued that this Object-Based Model represents a form of object-based auditory attention since it acts on an abstracted representation of the stimulus, modifying the Fourier modulation profile separately from its phase spectrum. We finally showed how this model could be applied to a speech activity detection problem, an important first step in a number of automated speech processing tasks.

Overall, the results in this thesis represent an improvement of our understanding of the computational strategies at work in the central auditory system, and we have showed how study of such questions can inspire biomimetic signal processing techniques for practical automated speech processing tasks.

5.2 Future work

5.2.1 Beyond the sustained firing model

The framework presented in Chap. 2 can be extended in a number of ways. For instance, to address the linearity limitation of the STRF in general, it is worthwhile to consider a model based on a linear-nonlinear cascade [187]. As mentioned earlier, the auditory pathway is necessarily hierarchical, and warrants consideration of hierarchical computational models. Indeed, recent physiological evidence also indicates that the representation becomes increasingly complex and nonlinear as one moves from away thalamo-recipient layers in primary auditory cortex (for a review, see [4]).

Also, a recent computational study in vision by Cadieu and Olshausen [188] proposes a hierarchical generative model that explicitly unifies notions of sparse coding and temporal stability. In particular, a two-layer network learns a sparse input representation whose activations vary smoothly over time, whereas a second layer modulates the plasticity of the first layer, resulting in a smooth time-varying basis for image sequences. One can imagine that such a framework could be

extended to spectro-temporal acoustic stimuli.

5.2.2 Extending top-down, attentional framework

The discriminative framework presented in Chap. 3 can be extended in a number of ways. First, instead of varying the shapes of the raw STRFs (i.e., each time-frequency or modulation profile bin), it may be advantageous to adapt parametric representations of STRF processing based on Gabor filters. Since the optimization considered in this study takes place over tens of thousands of parameters, adapting a simpler representation that contains far fewer parameters will enable the framework to scale to large data sets and more complex tasks.

Second, because auditory scene analysis generally involves complex sounds mixtures involving many sound classes, it is also of interest to consider STRF ensemble adaptation for discrimination problems beyond two categories. The linear discriminative model considered here was attractive largely because of its interpretable results, but extensions to multiple classes can be achieved using multi-class logistic regression or nonlinear multi-layered perceptrons. However, it remains to be seen whether induced plasticity in these settings would be consistent with the contrast filtering hypothesis and to what extent model predictions would correspond to neurophysiological results, which are unavailable to the best of our knowledge.

Third, it is also likely that the discriminability heuristic considered here is only part of the overall strategy by the auditory system to yield noise-robust representations of sound. Representation within primary auditory areas (and beyond) seem to be inherently noise robust, so it is therefore of interest to explore the impact of introducing a robustness term into the objective function ([169, 171]).

Finally, the results of Chap. 4 motivate further investigation other practical speech tasks. For instance, because the Object-Based model involves perturbation of a large number of parameters and consequently involves a good deal of computation, it is of interest to explore ways to approximate the modulation-domain adaptation process. One approach could be simply to perturb the STRF

ensemble modulation profile according to the average difference between the modulation profiles of the target and the reference, with multiplicative factors that control the relative influence of the target/reference. It is also of interest to explore how the Object-Based model can be used as a preprocessing step for automatic speech recognition and speaker verification tasks.

Appendix A

Supporting Information for Chapter 2

Supplemental Figure 1

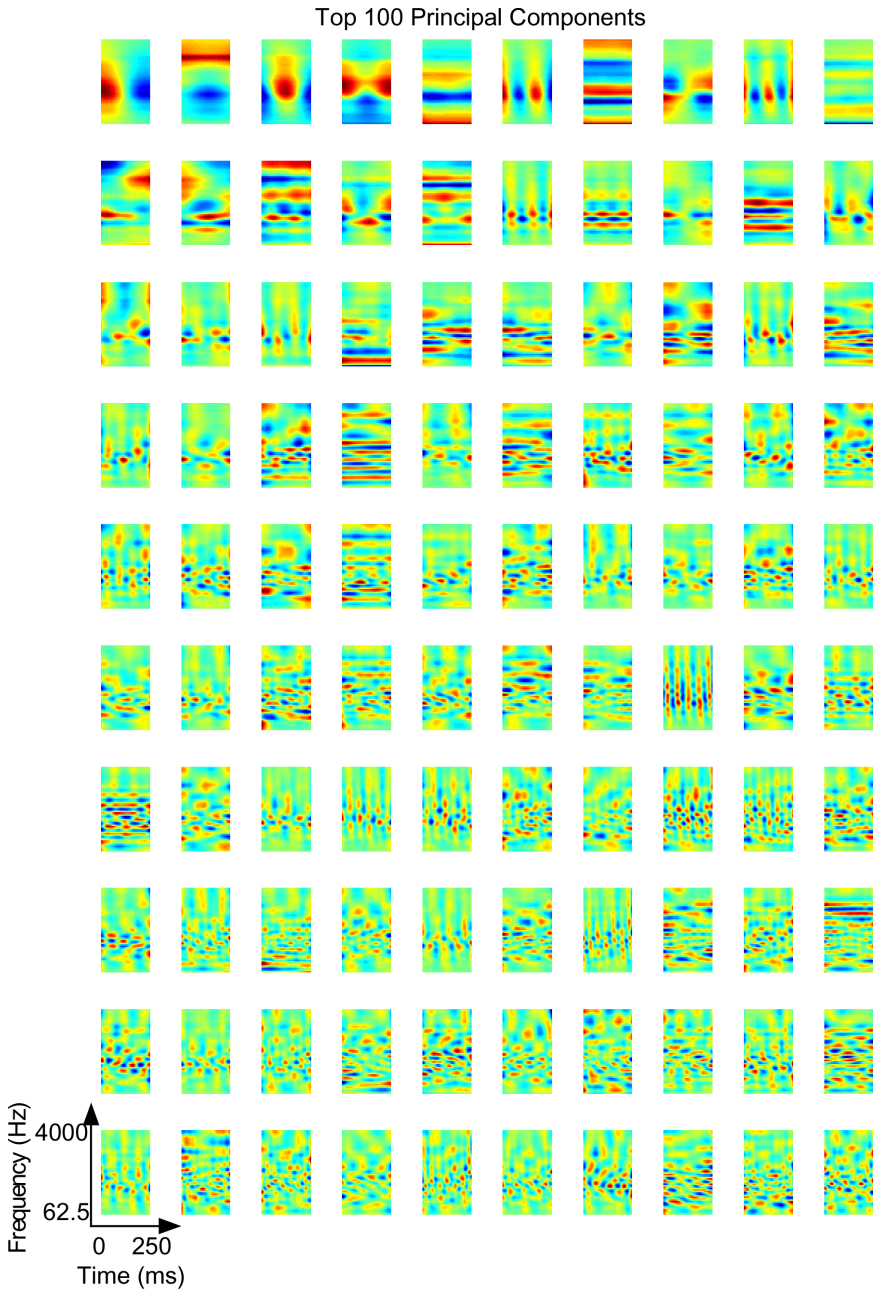

Figure A.1: Top 100 principal components of the natural stimulus ensemble.

Supplemental Figure 2

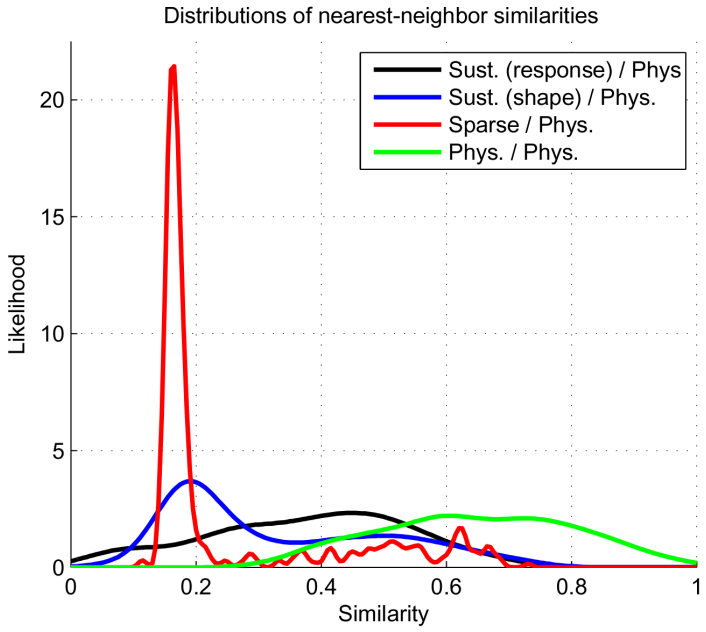

Figure A.2: STRFs corresponding to the top 10% "most persistent" responses for $\Delta T = 125$ ms.

Supplemental Figure 3

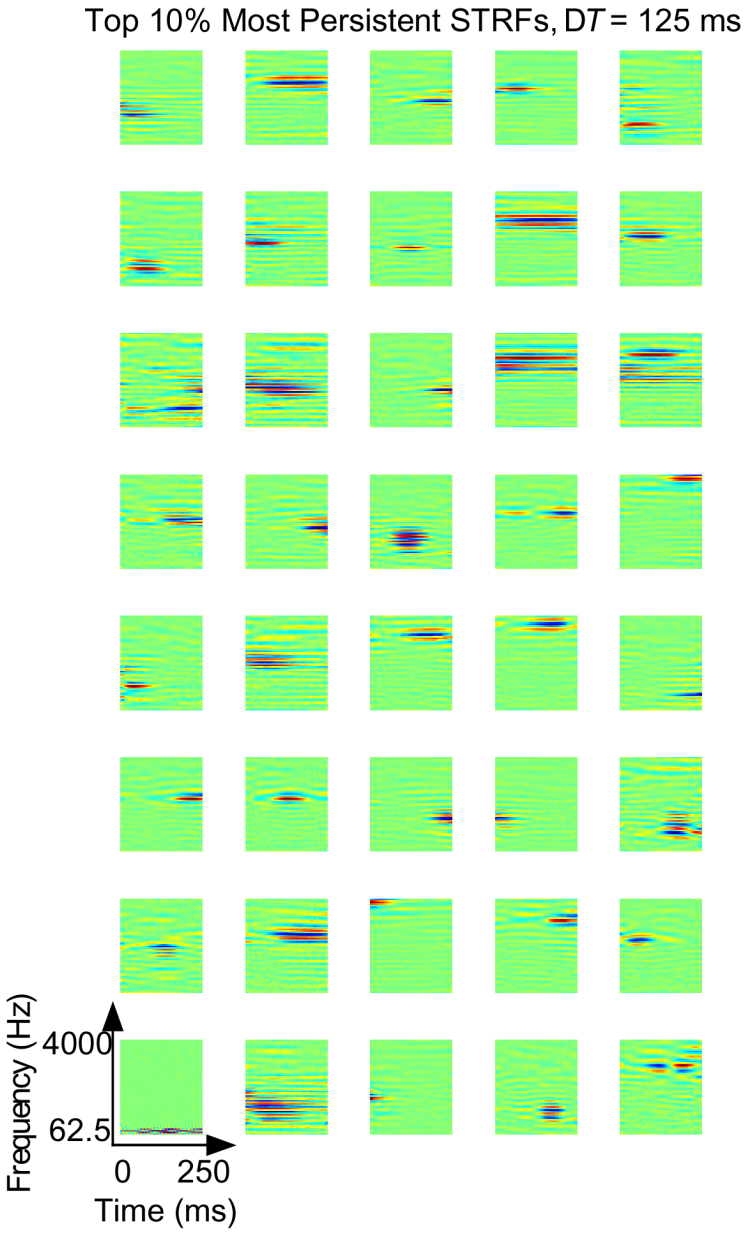

Figure A.3: Distributions of nearest-neighbor similarities for the model ensembles (response- and shape constrained sustained objective vs. the sparse objective) and the neural ensemble.

Appendix B

Supporting Information for

Chapter 3

Supplemental Figure 1

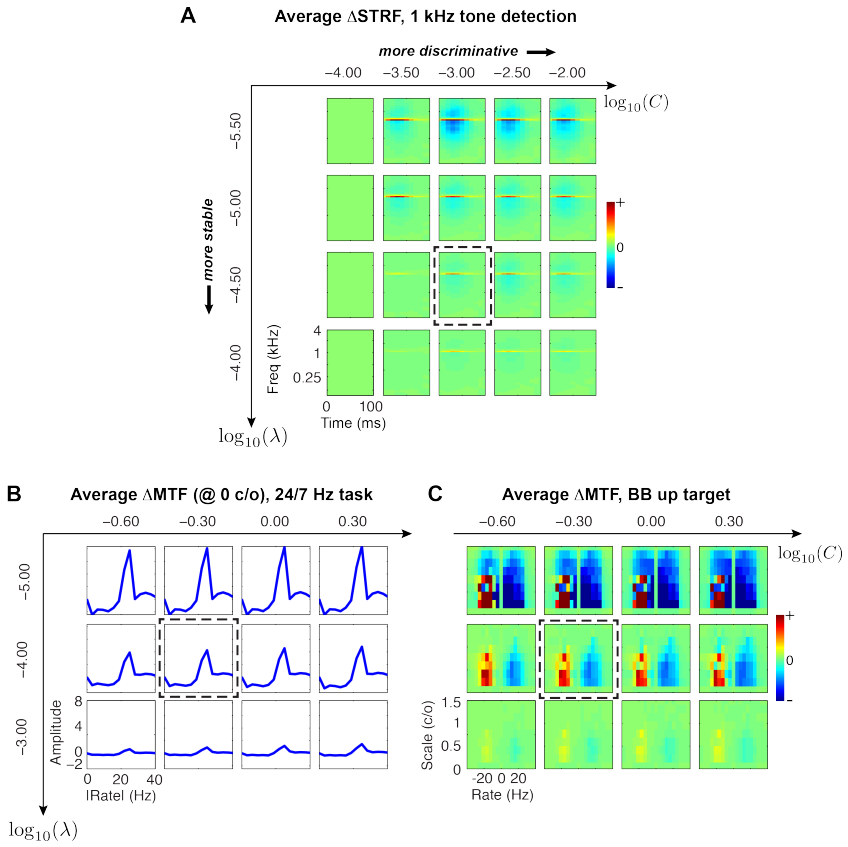

Figure B.1: Influence of model hyperparameters on population plasticity patterns. *Feature-Based Model, tone detection:* Panel A shows the average $\Delta STRF$ for the 1 kHz tone detection task as a function of model hyperparameters C and λ. Increasing C emphasizes discriminability in the overall objective function (Eq. 4 in the main text) whereas increasing λ emphasizes stability. Similar patterns were observed for the other tasks considered for the Feature-Based Model. *Object-Based Model, click rate discrimination:* Panel B shows the average ΔMTF at 0 c/o, folded along the scale axis for clarity. Changes to C and λ vary the magnitude of the plasticity effect as with the Feature-Based Model. Similar patterns were observed for the other tasks. *Object-Based Model, spectrotemporal modulation noise discrimination:* Finally, panel C shows the average ΔMTF for the *BB Up* target. Again, the size of the plasticity effect depends on choice of hyperparameters. Similar patterns were observed for the other tasks. For all panels, dashed boxes indicate the hyperparameters used for simulation and analysis for the various tasks described in the text.

Supplementary Text 1

Including a spectro-temporal mask in the Object-Based Model

As in the Feature-Based formulation, including a spectro-temporal mask in the Object-Based Model is desirable to guarantee locality of STRF adaptations. For technical reasons, however, use of a mask in the current formulation yields optimal STRF updates that fail to strictly satisfy the contrast filtering hypothesis, i.e., enhancement of target features and suppression of non-target features. As we show below, inclusion of the mask still induces adaptations that improve discrimination (by definition of the objective function), but it remains unclear whether this formulation of the model is suitable for accounting for physiological results. In this section, we elaborate on the details.

To begin, we model neural firing rate with the inclusion of a mask as

$$r_k(t,f;m) = [m_k(t,f) \cdot h_k^A(t,f)] *_{tf} s_m(t,f)$$

with Fourier domain representation

$$R_k(\omega,\Omega;m) = [M_k(\omega,\Omega) *_{\omega\Omega} H_k^A(\omega,\Omega)] \cdot S_m(\omega,\Omega)$$

where $M_k(\omega,\Omega), H_k(\omega,\Omega), S_m(\omega,\Omega)$ are the 2D Discrete Fourier Transforms of the mask, STRF, and stimulus, respectively. Expanding this term yields

$$\begin{aligned} R_k(\omega,\Omega;m) &= \sum_{ln} |M_k(\omega-l,\Omega-n)| \cdot |H_k^A(l,n)| \cdot |S_m(l,n)| \cdot \exp\{j\phi_{ln}(k,m)\} \\ &= \sum_{ln} |M_k(\omega-l,\Omega-n)| \cdot |H_k^A(l,n)| \cdot |S_m(l,n)| \cdot (\cos\phi_{ln}(k,m) + j\sin\phi_{ln}(k,m)) \\ &= \mathfrak{Re}\{R_k(\omega,\Omega;m)\} + j\mathfrak{Im}\{R_k(\omega,\Omega;m)\} \end{aligned}$$

where

$$\begin{aligned}
\phi_{ln}(k,m) &= \angle M_k(\omega-l,\Omega-n) + \angle H_k^0(l,n) + \angle S_m(l,n) \\
\mathfrak{Re}\{R_k(\omega,\Omega;m)\} &= \sum_{ln} |M_k(\omega-l,\Omega-n)| \cdot |H_k^A(l,n)| \cdot |S_m(l,n)| \cdot \cos\phi_{ln}(k,m) \\
\mathfrak{Im}\{R_k(\omega,\Omega;m)\} &= \sum_{ln} |M_k(\omega-l,\Omega-n)| \cdot |H_k^A(l,n)| \cdot |S_m(l,n)| \cdot \sin\phi_{ln}(k,m)
\end{aligned}$$

Note that because we optimize only the modulation profiles of the STRFs, we keep the phase of the STRFs fixed to those of the initial filters. The power spectrum of the neural response can therefore be expressed as

$$|R_k(\omega,\Omega;m)|^2 = \mathfrak{Re}^2\{R_k(\omega,\Omega;m)\} + \mathfrak{Im}^2\{R_k(\omega,\Omega;k)\}$$

We use the power spectrum in the objective function for mathematical convenience when computing gradients.

Next, consider the objective function defined as

$$J(\mathbf{w},\widehat{\mathcal{H}}_A) := \frac{1}{2}||\mathbf{w}||_2^2 - C\cdot\left\langle \log\sigma\left(y_m\left(w_0 + \sum_k w_k \sum_{\omega\Omega}|R_k(\omega,\Omega;m)|^2\right)\right)\right\rangle_m + \frac{\lambda}{2}\sum_k ||\Delta_k(\omega,\Omega)||_F^2$$

where $\widehat{\mathcal{H}}_A := \{|H_k(\omega,\Omega)|\}_{k=1}^K$ is the collection of ensemble modulation profiles and $\Delta_k(\omega,\Omega) := |H_k^0(\omega,\Omega)| - |H_k^A(\omega,\Omega)|$. As before, the optimal STRF modulation profiles are achieved by searching for a minimum of the of the objective function, i.e., when $\nabla_{|H_k^A(\omega,\Omega)|}J = 0$. Assuming this minimum occurs in the feasible set formed by the nonnegativity constraints on the regressor and STRF modulation profiles, we find

$$|H_k^A(\omega',\Omega')| = |H_k^0(\omega',\Omega')| + \frac{C}{\lambda}\cdot\left\langle y_m[1-\sigma(y_m\mathbf{w}^T\mathbf{R}_m)]\cdot\frac{\partial|R_k(\omega,\Omega)|^2}{\partial|H_k^A(\omega',\Omega')|}\right\rangle_m$$

where

$$\frac{\partial |R_k(\omega,\Omega)|^2}{\partial |H_k^A(\omega',\Omega')|} = 2 \cdot \mathfrak{Re}\{R_k(\omega,\Omega;m)\} \cdot [|M_k(\omega-\omega',\Omega-\Omega')| \cdot |S_m(\omega',\Omega')| \cdot \cos\phi_{\omega'\Omega'}(k,m)] + 2 \cdot \mathfrak{Im}\{R_k(\omega,\Omega;m)\} \cdot [|M_k(\omega-\omega',\Omega-\Omega')| \cdot |S_m(\omega',\Omega')| \cdot \sin\phi_{\omega'\Omega'}(k,m)]$$

for some particular (ω', Ω'). Recall that in the original Object-Based Model formulation, modulations associated with target stimuli were guaranteed to be enhanced whereas modulations associated with non-target stimuli were guaranteed to be suppressed. This was due to the nonnegativity of the filter response gradient $\partial|R_k(\omega,\Omega)|^2/\partial|H_k^A(\omega',\Omega')|$, and we therefore concluded that this model was strictly consistent with the contrast filtering hypothesis. However, observe that when the model includes a spectro-temporal mask, the filter response gradient involves sin and cos terms that necessarily take values in $[-1,+1]$. As a consequence, the gradient is not necessarily nonnegative and we are *no longer guaranteed that target modulations are enhanced and non-target modulations suppressed*. For this reason, this formulation of the model is *not* consistent with the contrast filtering hypothesis.

Under what conditions is the nonlinear formulation with a mask consistent with the contrast filtering hypothesis? Observe that when $0 \leq \phi_{\omega'\Omega'}(k,m) < \pi/2$, the sin and cos terms are positive and consequently the filter response gradient is also positive. However, unless we optimize the full Fourier domain representation, which gives access to the STRF phase but defeats the purpose of optimizing the modulation profiles of the STRFs, we have no control over whether this inequality is satisfied.

Bibliography

[1] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 3rd ed. Emerald Group Publishing Limited, 2008.

[2] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[3] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[4] T. O. Sharpee, C. A. Atencio, and C. E. Schreiner, "Hierarchical representations in the auditory cortex," *Current Opinion in Neurobiology*, vol. 21, no. 5, pp. 761–767, 2011.

[5] I. Nelken, "Processing of complex stimuli and natural scenes in the auditory cortex," *Current Opinion in Neurobiology*, vol. 14, no. 4, pp. 474–480, 2004.

[6] T. D. Griffiths and J. D. Warren, "What is an auditory object?" *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 887–892, 2004.

[7] J. K. Bizley and Y. E. Cohen, "The what, where and how of auditory-object perception," *Nature Reviews Neuroscience*, vol. 14, no. 10, pp. 693–707, 2013.

[8] C. A. Atencio and C. E. Schreiner, *Stimulus choices for spike-triggered receptive field analyses*, ser. Handbook of Modern Techniques in Auditory Cortex. Nova Science Pub Inc, 2013.

BIBLIOGRAPHY

[9] A. M. H. J. Aertsen and P. I. M. Johannesma, "Spectro-temporal receptive fields of auditory neurons in the grassfrog. I. Characterization of tonal and natural stimuli," *Biological Cybernetics*, vol. 38, pp. 223–234, 1980.

[10] ——, "The spectro-temporal receptive field," *Biological Cybernetics*, vol. 42, pp. 133–143, 1981.

[11] J. J. Eggermont, A. M. Aertsen, and P. I. Johannesma, "Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field," *Hearing Research*, vol. 10, no. 2, pp. 167–190, 1983.

[12] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 85, pp. 1220–1234, 2001.

[13] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design," *Journal of Computational Neuroscience*, vol. 9, no. 1, pp. 85–111, 2000.

[14] F. E. Theunissen, K. Sen, and A. J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *Journal of Neuroscience*, vol. 20, no. 6, pp. 2315–2331, 2000.

[15] D. J. Klein, J. Z. Simon, D. A. Depireux, and S. A. Shamma, "Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex," *Journal of Computational Neuroscience*, vol. 20, no. 2, pp. 111–136, 2006.

[16] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature Neuroscience*, vol. 6, no. 11, pp. 1216–1223, 2003.

[17] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks," *Journal of Neuroscience*, vol. 25, no. 33, pp. 7623–7635, 2005.

[18] S. Atiani, M. Elhilali, S. V. David, J. B. Fritz, and S. A. Shamma, "Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields," *Neuron*, vol. 61, no. 3, pp. 467–480, 2009.

[19] O. Schwartz, J. W. Pillow, N. C. Rust, and E. P. Simoncelli, "Spike-triggered neural characterization," *Journal of Vision*, vol. 6, no. 4, 2006.

[20] A. Calabrese, J. W. Schumacher, D. M. Schneider, L. Paninski, and S. M. N. Wooley, "A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds," *PLoS ONE*, vol. 6, no. 1, p. e16104, 2011.

[21] V. Volterra, *Theory of Functionals and Integral and Integro-Differential Equations.* New York: Dover Publications, 1959.

[22] M. Elhilali, S. A. Shamma, J. Z. Simon, and J. B. Fritz, *A Linear Systems View to the Concept of STRF*, ser. Handbook of Modern Techniques in Auditory Cortex. Nova Science Pub Inc, 2013.

[23] M. Sahani and J. F. Linden, "How linear are auditory cortical responses?" in *Adv. Neural Inf. Proc. Sys. (NIPS)*, 2002.

[24] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," *Sensory Comm.*, pp. 217–234, 1961.

[25] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, 2001.

[26] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Curr. Op. Neurobio.*, vol. 14, pp. 481–487, 2004.

[27] S. B. Laughlin, "Energy as a constraint on the coding and processing of sensory information," *Curr. Op. Neurobio.*, vol. 11, no. 4, pp. 475–480, 2001.

[28] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[29] W. E. Vinje and J. L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[30] M. R. DeWeese, M. Wehr, and A. M. Zador, "Binary spiking in auditory cortex," *J. Neurosci.*, vol. 23, no. 21, pp. 7940–7949, 2003.

[31] T. Hromádka, M. R. DeWeese, and A. M. Zador, "Sparse representation of sounds in the unanesthetized auditory cortex," *PLoS Bio.*, vol. 6, no. 1, p. e16, 2008.

[32] D. J. Klein, P. König, and K. P. Körding, "Sparse spectrotemporal coding of sounds," *EURASIP J. Appl. Sig. Proc.*, vol. 2003, no. 7, pp. 659–667, 2003.

[33] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, 2006.

[34] N. L. Carlson, V. L. Ming, and M. R. DeWeese, "Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus," *PLoS Comp. Bio.*, vol. 8, no. 7, p. e1002594, 2012.

[35] X. Wang, T. Lu, R. K. Snider, and L. Liang, "Sustained firing in auditory cortex evoked by preferred stimuli," *Nature*, vol. 435, pp. 341–346, 2005.

[36] X. Wang, T. Lu, D. Bendor, and E. Bartlett, "Neural coding of temporal information in auditory thalamus and cortex," *Neuroscience*, vol. 157, pp. 484–493, 2008.

[37] C. I. Petkov, K. N. O'Connor, and M. L. Sutter, "Encoding of illusory continuity in primary auditory cortex," *Neuron*, vol. 54, pp. 153–165, 2007.

[38] X. Wang, "Neural coding strategies in auditory cortex," *Hearing Research*, vol. 229, pp. 81–93, 2007.

[39] M. Elhilali, J. B. Fritz, D. J. Klein, J. Z. Simon, and S. A. Shamma, "Dynamics of precise spike timing in primary auditory cortex." *Journal of Neuroscience*, vol. 24, no. 5, pp. 1159–1172, 2004.

[40] M. Elhilali and S. A. Shamma, "A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation." *Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3751–3771, 2008.

[41] J. Hurri and A. Hyvärinen, "Simple-cell-like receptive fields maximize temporal coherence in natural video," *Neural Comp.*, vol. 15, pp. 663–691, 2003.

[42] K. P. Körding, C. Kayser, W. Einhäuser, and P. König, "How are complex cell properties adapted to the statistics of natural stimuli?" *J. Neurophys.*, vol. 91, pp. 206–212, 2004.

[43] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: unsupervised learning of invariances," *Neural Comp.*, vol. 14, pp. 715–770, 2002.

[44] J. C. Middlebrooks, "Auditory cortex cheers the overture and listens through the finale," *Nature Neurosci.*, vol. 8, no. 7, pp. 851–852, 2005.

[45] C. D. Gilbert and W. Li, "Adult visual cortical plasticity," *Neuron*, vol. 75, no. 2, pp. 250–264, 2012.

[46] D. E. Feldman and M. Brecht, "Map plasticity in somatosensory cortex," *Science*, vol. 310, no. 5749, pp. 810–815, 2005.

[47] N. Mandairon and C. Linster, "Odor perception and olfactory bulb plasticity in adult mammals," *Journal of Neurophysiology*, vol. 101, no. 5, pp. 2204–2209, 2009.

[48] C. E. Schreiner and D. B. Polley, "Auditory map plasticity: diversity in causes and consequences," *Current Opinion in Neurobiology*, vol. 24, no. 0, pp. 143–156, 2014.

[49] B. C. Motter, "Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli," *Journal of Neurophysiology*, vol. 70, no. 3, pp. 909–919, 1993.

[50] S. V. David, B. Y. Hayden, J. A. Mazer, and J. L. Gallant, "Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision," *Neuron*, vol. 59, no. 3, pp. 509–521, 2008.

[51] J. Ahveninen, M. Hamalainen, I. P. Jaaskelainen, S. P. Ahlfors, S. Huang, F. H. Lin, T. Raij, M. Sams, C. E. Vasios, and J. W. Belliveau, "Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise," *Proceedings of the National Academy of Sciences*, vol. 108, no. 10, pp. 4182–4187, 2011.

[52] S. Atiani, S. V. David, D. Elgueda, M. Locastro, S. Radtke-Schuller, S. A. Shamma, and J. B. Fritz, "Emergent selectivity for task-relevant stimuli in higher-order auditory cortex," *Neuron*, vol. 82, no. 2, pp. 486–499, 2014.

[53] A. Treisman, "The binding problem," *Current Opinion in Neurobiology*, vol. 6, no. 2, pp. 171–178, 1996.

[54] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[55] L. Shuai and M. Elhilali, "Task-dependent neural representations of salient events in dynamic auditory scenes," *Frontiers in Neuroscience*, vol. 8, no. 203, 2014.

[56] E. K. Miller and T. J. Buschman, "Cortical circuits for the control of attention," *Current Opinion in Neurobiology*, vol. 23, no. 2, pp. 216–222, 2013.

[57] M. Carrasco, "Visual attention: the past 25 years," *Vision Research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[58] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[59] L. Itti, G. Rees, and J. K. Tsotsos, Eds., *Neurobiology of Attention*. Academic Press, 2005.

[60] S. Treue and J. C. M. Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature*, vol. 399, no. 6736, pp. 575–579, 1999.

[61] H. Spitzer, R. Desimone, and J. Moran, "Increased attention enhances both behavioral and neuronal performance," *Science*, vol. 240, no. 4850, pp. 338–340, 1988.

[62] T. Womelsdorf, K. Anton-Erxleben, F. Pieper, and S. Treue, "Dynamic shifts of visual receptive fields in cortical area MT by spatial attention," *Nature Neuroscience*, vol. 9, no. 9, pp. 1156–1160, 2006.

[63] J. Martinez-Trujillo and S. Treue, "Attentional modulation strength in cortical area MT depends on stimulus contrast," *Neuron*, vol. 35, no. 2, pp. 365–370, 2002.

[64] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations," *ACM Transactions on Applied Perception*, vol. 7, no. 1, pp. 1–39, 2010.

[65] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[66] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *Journal of Neuroscience*, vol. 13, no. 11, pp. 4700–19, 1993.

[67] J. Tsotsos, S. Culhane, W. K. Wai, and Y. Lai, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 3702, no. 95, 1995.

[68] V. Navalpakkam and L. Itti, "Search goal tunes visual features optimally," *Neuron*, vol. 53, no. 4, pp. 605–617, 2007.

[69] J. H. Reynolds and D. J. Heeger, "The normalization model of attention," *Neuron*, vol. 61, no. 2, pp. 168–85, 2009.

[70] A. Borji and L. Itti, "Optimal attentional modulation of a neural population." *Frontiers in Computational Neuroscience*, vol. 8, no. March, p. 34, 2014.

[71] F. Baluch and L. Itti, "Mechanisms of top-down attention," *Trends in Neurosciences*, vol. 34, no. 4, pp. 210–224, 2011.

[72] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1?" *Hearing Research*, vol. 229, no. 1-2, pp. 186–203, 2007.

[73] ——, "Auditory attention–focusing the searchlight on sound," *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[74] V. M. Bajo and A. J. King, "Focusing attention on sound," *Nature Neuroscience*, vol. 13, no. 8, pp. 913–914, 2010.

[75] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Adaptive changes in cortical receptive fields induced by attention to complex sounds," *Journal of Neurophysiology*, vol. 98, no. 4, pp. 2337–2346, 2007.

[76] S. V. David, J. B. Fritz, and S. A. Shamma, "Task reward structure shapes rapid receptive field plasticity in auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 109, no. 6, pp. 2144–2149, 2012.

[77] P. Yin, J. B. Fritz, and S. A. Shamma, "Rapid spectrotemporal plasticity in primary auditory cortex during behavior," *The Journal of Neuroscience*, vol. 34, no. 12, pp. 4396–4408, 2014.

[78] M. Elhilali, J. B. Fritz, T. S. Chi, and S. A. Shamma, "Auditory cortical receptive fields: stable entities with plastic abilities," *Journal of Neuroscience*, vol. 27, no. 39, pp. 10 372–10 382, 2007.

[79] J. B. Fritz, S. V. David, S. Radtke-Schuller, P. Yin, and S. A. Shamma, "Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex," *Nature Neuroscience*, vol. 13, no. 8, pp. 1011–1019, 2010.

[80] N. Mesgarani, J. Fritz, and S. Shamma, "A computational model of rapid task-related plasticity of auditory cortical receptive fields," *Journal of Computational Neuroscience*, vol. 28, no. 1, pp. 19–27, 2010.

[81] T. D. Griffiths, J. D. Warren, S. K. Scott, I. Nelken, and A. J. King, "Cortical processing of complex sound: a way forward?" *Trends in Neurosciences*, vol. 27, no. 4, pp. 181–185, 2004.

[82] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus.* Linguistic Data Consortium, Philadelphia, 1993.

[83] *The BBC Sound Effects Library Original Series.* http://www.soundideas.com, 2006.

[84] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.

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

[86] J. B. Rosen, "The gradient projection method for nonlinear programming: part II. nonlinear constraints." *J. Soc. Indust. Appl. Math.*, vol. 9, no. 4, pp. 514–532, 1961.

[87] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

[88] B. K. P. Horn, H. M. Hilden, and S. Negahdaripour, "Closed form solution of absolute orientation using orthonormal matrices," *J. Optical Soc. A*, vol. 5, no. 7, pp. 1127–1135, 1988.

[89] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge Univ. Press, 1985.

[90] L. M. Miller, M. A. Escabí, H. L. Read, and C. E. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," *J. Neurophys.*, vol. 87, pp. 516–527, 2002.

[91] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Adv. Neural Inf. Proc. Sys. (NIPS)*, 2002.

[92] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design," *Journal of Computational Neuroscience*, vol. 9, no. 1, pp. 85–111, 2000.

[93] S. V. David, N. Mesgarani, J. B. Fritz, and S. A. Shamma, "Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli," *Journal of Neuroscience*, vol. 29, no. 11, pp. 3374–3386, 2009.

[94] B. Willmore and D. J. Tolhurst, "Characterizing the sparseness of neural codes," *Network: Computation in Neural Systems*, vol. 12, pp. 255–270, 2001.

[95] S. V. David, N. Mesgarani, J. B. Fritz, and S. A. Shamma, "Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli," *J. Neurosci.*, vol. 29, no. 11, pp. 3374–3386, 2009.

[96] S. M. N. Wooley, P. R. Gill, T. Fremouw, and F. E. Theunissen, "Functional groups in the avian auditory system," *J. Neurosci.*, vol. 20, no. 9, pp. 2780–2793, 2009.

[97] S. Rosen, "Temporal information in speech: acoustic, auditory, and linguistic aspects," *Phil. Trans. R. Soc. Lond. B*, vol. 336, pp. 367–373, 1992.

[98] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *J. Acoust. Soc. Am.*, vol. 114, no. 6, pp. 3394–3411, 2003.

[99] P. Berkes and L. Wiskott, "Slow feature analysis yields a rich repertoire of complex cell properties," *J. Vision*, vol. 5, pp. 579–602, 2005.

[100] H. Attias and C. E. Schreiner, "Temporal low-order statistics of natural sounds," in *Adv. Neural Inf. Proc. Sys. (NIPS)*, 1997.

[101] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[102] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Computational Biology*, vol. 5, no. 3, p. e1000302, 2009.

[103] S. K. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 21, no. 2, pp. 416–426, 2013.

[104] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comp. Bio.*, vol. 5, no. 3, p. e1000302, 2009.

[105] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic, Dordrecht, 1994.

[106] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 37, pp. 1641–1648, 1989.

[107] J. Pinto, S. Garimella, M. Magimai-Doss, H. Hermansky, and H. Bourlard, "Analyzing MLP-based hierarchical phoneme posterior probability estimator," *IEEE Trans. Speech and Audio Process*, vol. 19, pp. 225–241, 2011.

[108] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Speech Research Unit, Defense Research Agency, Malvern, U.K., Tech. Rep., 1992.

[109] H. Hirsch. (2005) FaNT: Filtering and noise adding tool.

[110] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 257–270, 2007.

[111] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 2, pp. 382–395, 1994.

[112] W. Hashimoto, "Quadratic forms in natural images," *Network: Computation in Neural Systems*, vol. 14, pp. 765–788, 2003.

[113] S. M. N. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," *Nature Neurosci.*, vol. 8, no. 10, pp. 1371–1379, 2005.

[114] K. I. Nagel and A. J. Doupe, "Organizing principles of spectro-temporal encoding in the avian primary auditory area Field L," *Neuron*, vol. 58, pp. 938–955, 2008.

[115] F. A. Rodriguez, C. Chen, H. L. Read, and M. A. Escabí, "Neural modulation tuning characteristics scale to efficiently encode natural sound statistics," *J. Neurosci.*, vol. 30, no. 47, pp. 15 969–15 980, 2010.

[116] P. Földiák, "Learning invariances from transformational sequences," *Neural Comp.*, vol. 3, pp. 194–2000, 1991.

[117] G. Mitchison, "Removing time variation with the anti-Hebbian differential synapse," *Neural Comp.*, vol. 3, no. 312–320, 1991.

[118] S. Becker, "Learning to categorize objects using temporal coherence," in *Adv. Neural Inf. Proc. Sys. (NIPS)*, 1993, pp. 361–368.

[119] A. Hyvärinen, "Blind source separation by nonstationarity of variance: a cumulant-based approach," *IEEE Trans. Neural Networks*, vol. 12, no. 6, pp. 1471–1474, 2001.

[120] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[121] P. Berkes and L. Wiskott, "On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields," *Neural Comp.*, vol. 18, no. 8, pp. 1868–1895, 2006.

[122] T. Blaschke, P. Berkes, and L. Wiskott, "What is the relation between slow feature analysis and independent component analysis?" *Neural Comp.*, vol. 18, no. 10, pp. 2495–2508, 2006.

[123] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, ser. Wiley series on adaptive and learning systems for signal processing, communications, and control. John Wiley and Sons, 2001.

[124] S. Klampfl and W. Maass, "A theoretical basis for emergent pattern discrimination in neural systems through slow feature extraction," *Neural Comp.*, vol. 22, pp. 2979–3035, 2010.

[125] F. Creutzig and H. Sprekeler, "Predictive coding and the slowness principle: an information-theoretic approach," *Neural Comp.*, vol. 20, no. 4, pp. 1026–1041, 2008.

[126] H. Sprekeler, "On the relation of slow feature analysis and Laplacian eigenmaps," *Neural Comp.*, vol. 23, pp. 3287–3302, 2011.

[127] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 2, pp. 382–395, 1994.

[128] S. K. Nemala, K. Patil, and M. Elhilali, "Multistream bandpass modulation features for robust speech recognition," in *Interspeech*, 2011.

[129] S. K. Nemala, "Robust Speech Recognition by Humans and Machines: The Role of Spectro-Temporal Modulations," Ph.D. dissertation, Johns Hopkins University, 2012.

[130] M. A. Carlin and M. Elhilali, "Sustained firing of central auditory neurons yields a discriminative spectro-temporal representation for natural sounds," *PLoS Computational Biology*, vol. 9, no. 3, p. e1002982, 2013.

[131] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[132] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.

[133] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[134] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[135] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[136] C. Alain and S. R. Arnott, "Selectively attending to auditory objects," *Frontiers in Bioscience*, vol. 5, pp. D202–12, 2000.

[137] K. Krumbholz, S. B. Eickhoff, and G. R. Fink, "Feature- and object-based attentional modulation in the human auditory 'where' pathway," *Journal of Cognitive Neuroscience*, vol. 19, no. 10, pp. 1721–1733, 2007.

[138] R. E. Beitel, C. E. Schreiner, S. W. Cheung, X. Wang, and M. M. Merzenich, "Reward-dependent plasticity in the primary auditory cortex of adult monkeys trained to discriminate temporally modulated signals," *Proceedings of the National Academy of Sciences*, vol. 100, no. 19, pp. 11 070–11 075, 2003.

[139] J. Fritz, M. Elhilali, P. Yin, N. Harper, K. Donaldson, and S. A. Shamma, "Multiple auditory tasks and the single cortical neuron: salient temporal and spectral cues drive orthogonal, dynamic, task-related receptive plasticity in primary auditory cortex," in *Society for Neuroscience Meeting, Washington, DC*, 2005 2005.

[140] J. Fritz, M. Elhilali, and S. Shamma, "Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex," *Hearing Research*, vol. 206, no. 1-2, pp. 159–176, 2005.

[141] M. P. Kilgard and M. M. Merzenich, "Plasticity of temporal information processing in the primary auditory cortex." *Nature Neuroscience*, vol. 1, no. 8, pp. 727–31, 1998.

[142] M. P. Kilgard, P. K. Pandya, J. Vazquez, A. Gehi, C. E. Schreiner, and M. M. Merzenich, "Sensory input directs spatial and temporal plasticity in primary auditory cortex," *Journal of Neurophysiology*, vol. 86, no. 1, pp. 326–338, 2001.

[143] D. Rasmusson, S. Smith, and K. Semba, "Inactivation of prefrontal cortex abolishes cortical acetylcholine release evoked by sensory or sensory pathway stimulation in the rat," *Neuroscience*, vol. 149, no. 1, pp. 232–241, 2007.

[144] S. Shamma, J. Fritz, S. David, D. Winkowski, P. Yin, and M. Elhilali, *Correlates of auditory attention and task performance in primary auditory and prefrontal cortex*, ser. The Neurophysiological Bases of Auditory Perception. New York: Springer, 2010, ch. 51, pp. 555–570.

[145] S. Bao, V. T. Chan, and M. M. Merzenich, "Cortical remodelling induced by activity of ventral tegmental dopamine neurons," *Nature*, vol. 412, no. 6842, pp. 79–83, 2001.

[146] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, pp. 331–348, 2003.

[147] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: the biological bases of musical timbre perception," *PLoS Computational Biology*, vol. 8, no. 11, p. e1002759, 2012.

[148] J. B. Fritz, S. David, and S. Shamma, *Attention and Dynamic, Task-Related Receptive Field Plasticity in Adult Auditory Cortex*, ser. Springer Handbook of Auditory Research. New York, NY: Springer New York, 2013, ch. 45.

[149] S. Greenberg, A. Popper, and W. Ainsworth, *Speech Processing in the Auditory System*. Springer, Berlin, 2004.

[150] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 14, no. 3, pp. 920–930, 2006.

[151] N. Mesgarani and S. Shamma, "Denoising in the domain of spectro-temporal modulations," *EURASIP Journal on Audio, Speech, and Music Processing*, p. 042357, 2007.

[152] M. A. Carlin and M. Elhilali, "Modeling attention-driven plasticity in auditory cortical receptive fields," *Frontiers in Computational Neuroscience (under review)*, 2015.

[153] L. M. Miller, M. A. Escabi, H. L. Read, and C. E. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," *Journal of Neurophysiology*, vol. 87, no. 1, pp. 516–527, 2002.

[154] M. A. Escabi and H. L. Read, "Representation of spectrotemporal sound information in the ascending auditory pathway," *Biological Cybernetics*, vol. 89, no. 5, pp. 350–362, 2003.

[155] A. Benyassine, E. Shlomot, H. yu Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *Communications Magazine, IEEE*, vol. 35, no. 9, pp. 64–73, 1997.

[156] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for german connected digit recognition." in *EUROSPEECH*, vol. 99, 1999, pp. 61–64.

[157] K.-H. Woo, T.-Y. Yang, K.-J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *IEEE Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[158] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselỳ, and P. Matejka, "Developing a Speech Activity Detection System for the DARPA RATS Program," in *Interspeech*, 2012.

[159] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *ICASSP*, 2014, pp. 2519–2523.

[160] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[161] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, 1997, pp. 1895–1898.

[162] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech*, 2010.

[163] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *ICASSP*, 2010, pp. 4466–4469.

[164] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland, "Reading a neural code." *Science*, vol. 252, no. 5014, pp. 1854–1857, Jun 1991.

[165] R. R. de Ruyter van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek, "Reproducibility and variability in neural spike trains." *Science*, vol. 275, no. 5307, pp. 1805–1808, Mar 1997.

[166] G. T. Buracas, A. M. Zador, M. R. DeWeese, and T. D. Albright, "Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex." *Neuron*, vol. 20, no. 5, pp. 959–969, May 1998.

[167] Y. Miyawaki, H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders." *Neuron*, vol. 60, no. 5, pp. 915–929, Dec 2008.

[168] N. Mesgarani, S. David, and S. Shamma, "Influence of context and behavior on the population code in primary auditory cortex," *Journal of Neurophysiology*, vol. 102, pp. 3329–3333, 2009.

[169] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Mechanisms of noise robust representation of speech in primary auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 18, pp. 6792–6797, 2014.

[170] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS Biol*, vol. 10, no. 1, p. e1001251, 2012.

[171] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[172] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[173] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, 2014.

[174] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech

quality assessment of narrow-band telephone networks and speech codecs." ITU-T P.862, Tech. Rep., 2001.

[175] S. Nemala, D. Zotkin, R. Duraiswami, and M. Elhilali, "Biomimetic multi-resolution analysis for robust speaker recognition," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 22, 2012.

[176] M. A. Carlin, K. Patil, S. K. Nemala, and M. Elhilali, "Robust phoneme recognition using biomimetic speech contours," in *Interspeech*, 2012.

[177] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans.Speech and Audio Process.*, vol. 2, no. 4, pp. 382–395, 1994.

[178] S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," in *EUROSPEECH*, 1997.

[179] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[180] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[181] J. Ramirez, J. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.

[182] H. Soltau, G. Saon, and T. N. Sainath, "Joint training of convolutional and non-convolutional neural networks," in *ICASSP*, 2014.

[183] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 53, no. 5, pp. 753–767, 2011.

[184] K. Patil and M. Elhilali, "Goal-oriented auditory scene recognition," in *Interspeech*, 2012.

[185] ——, "Task-driven attentional mechanisms for auditory scene recognition," in *ICASSP*, 2013, pp. 828–832.

[186] A. Bellur and M. Elhilali, "Detection of Speech Tokens in Noise Using Adaptive Spectrotemporal Receptive Fields," in *49th Annual Conference on Information Sciences and Systems (CISS)*, 2015.

[187] A. Calabrese, J. W. Schumacher, D. M. Schneider, L. Paninski, and S. M. N. Wooley, "A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds," *PLoS ONE*, vol. 6, no. 1, p. e16104, 2011.

[188] C. F. Cadieu and B. A. Olshausen, "Learning transformational invariants from natural movies," in *Adv. Neural Inf. Proc. Sys. (NIPS)*, 2009.

Vita

Michael A. Carlin (b. 1982) received the B.S.E. and M.S.E. degrees in Electrical and Computer Engineering from Temple University in 2005 and 2007, respectively, and the M.S.E. degree in Electrical and Computer Engineering from Johns Hopkins University in 2011. Prior to enrolling in the Ph.D. program in Electrical and Computer Engineering at Johns Hopkins, he was a member of the technical staff at the Air Force Research Laboratory in Rome, NY from 2006–2008. While a graduate student, he participated in two SCALE workshops (2009 and 2013) at the Human Language Technology Center of Excellence, and interned at MIT Lincoln Laboratory in 2011. His research focuses on the computational principles that underlie representation, plasticity, and noise robustness in the auditory system, and how to apply this knowledge to automated sound processing systems. Michael was inducted into Eta Kappa Nu in 2004 and is a member of the IEEE.

Starting in March 2015, Michael will work as a Senior Data Scientist with RedOwl Analytics in Baltimore, MD.