

# **Single-unit representations of natural sound mixtures in auditory cortex**

By

Gregory R. Hamersky

A DISSERTATION

Presented to the Neuroscience Graduate Program  
and the Oregon Health & Science University School of Medicine

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

09/18/2024

**School of Medicine**  
Oregon Health & Science University

---

**CERTIFICATE OF APPROVAL**

---

This is to certify that the Ph.D. dissertation of  
GREGORY R. HAMERSKY  
has been approved on XX/XX/XXXX

---

Advisor, Stephen David

---

Member & Chair, Lina Reiss

---

Member, Frederick Gallun

---

Member, Vincent Costa

---

External member, Elizabeth Moss

# Contents

Acknowledgements.....	ii
Abstract.....	iii
1. Introduction.....	1
1.1    How do you make sense of sound in a busy acoustic world?.....	1
1.2    How does your brain make sense of sound in a busy acoustic world? .....	9
1.2.1    Anatomy and physiology of the central auditory system: from cochlea to cortex .....	9
1.2.2    Types of sound stimuli .....	12
1.2.3    Population level and psychoacoustic studies of auditory stream formation in humans .....	14
1.2.4    The role of animal models in studies of auditory streaming .....	23
1.2.5    Single-unit representations and auditory coding .....	29
1.2.6    How does my work fit into this context? .....	35
2. Single-unit representations of natural background/foreground contrasts in passively listening ferrets..	37
Abstract.....	38
Significance statement .....	39
Introduction.....	40
Material and Methods .....	41
Results.....	47
Discussion.....	65
3. Single-unit representations of background/foreground contrasts in trained, behaving ferrets.....	69
Abstract.....	70
Introduction.....	71
Methods .....	72
Results.....	75
Discussion.....	82
4. Conclusions and future directions.....	86
4.1    Natural spectrotemporal statistics play a crucial role in the interactions of foreground and background sounds.....	87
4.2    The role of binaural cues in natural sound representations .....	89
4.3    Contributions to foreground response reduction outside of AC? .....	90
4.4    Behavioral consequences of foreground response reduction.....	91
4.5    Modeling complex representations of natural sounds using deep learning .....	95
References.....	99

## **Acknowledgements**

The last six years have been filled with outdoor merriment, climbing walls, pinballing to exhaustion, biking in various levels of dress, frisbees, sitting in circles in the park, sitting in circles in the park throwing darts at one another, pretzels, gathering wood, hoarding wood, turning wood, profound and persistent feelings of dread, spicy air, watering plants, birdwatching, treewatching, gin and no juice, and a good amount of shenanigans. How lucky I am to have found such a wonderful assortment of people to experience all of this with. I am not the person I showed up here as because of them, and that couldn't please me more. Thanks to everyone for everything.

## **Abstract**

The perception of important, behaviorally salient sounds in a world cluttered with competing background noise requires the ability to segregate relevant from irrelevant sound sources. It remains unknown exactly how and at what level of auditory processing neural representations of complex and noisy auditory scenes are refined to allow for noise-robust perception. Much of the prior work investigating neural mechanisms of this contrast between behaviorally meaningful sounds and background noise has supported the theory that auditory cortex activity is largely invariant to noise. While these results are consistent with behavioral studies and phenomenological evidence that emphasize the perception of behaviorally salient sounds, it remains unclear what information about background noise is represented at the single-unit level. In this dissertation, I take a step towards a more complete understanding of how complex auditory scenes—both the important foregrounds and the noisy backgrounds—of natural sounds are represented at the single-unit level in early stages of auditory processing. I will present results that challenge intuitions about the ubiquitousness of robust noise-invariant representation of behaviorally salient stimuli across the auditory hierarchy. Here, contrary to prevailing theories, I show the preferential reduction of single-unit responses to natural foreground sounds when paired with natural background noise and present potential mechanisms by which this unexpected finding arises. These results support alternate hypotheses to explain the emergence of background invariance, such as the computation of enhanced representation of foreground-like sounds at later processing stages.

# 1. Introduction

In a two-part introduction, I will orient the reader to the background necessary to understand the presentation of my main results and, more importantly, to gain context for the motivation behind the questions to which I've spent several years trying to answer. The study of systems—specifically of sensory systems—in neuroscience is particularly amenable to a birds-eye overview thanks to broad accessibility afforded by its connections to and descriptions of daily experiences that can resonate with both technical and non-technical audiences.

With that in mind, the first section acknowledges the potential for diverse readership. Using non-technical descriptions of basic auditory neuroscience principles and relevant auditory phenomena, I first lay a broad foundation to gain an intuitive appreciation of the work's conceptual motivation. Once oriented, the second section more technically provides a context for where the more focused questions I ask reside within the existing literature, setting the stage for the subsequent descriptions and details of my contributions to science.

## 1.1 How do you make sense of sound in a busy acoustic world?

In everyday hearing, listeners are faced with a world cluttered with sound. Whether walking down a city street or a remote backcountry trail, silence is now more elusive than ever, as noise—be it agitated traffic congestion or chipper birdsong—vies for the listener's attention. Despite these distractions, most listeners have a remarkable yet effortless ability, called auditory streaming, to separate even extremely busy soundscapes into their distinct sources (Bregman, 1990). Moreover, even within these noisy worlds, most listeners can easily direct their attention to an important sound or rapidly and precisely shift their focus between numerous sound sources.

In psychology, the phenomenon of auditory streaming is classically described in the context of the cocktail-party effect, describing the ability of a partygoer to maintain a conversation with a partner in a room filled with countless other voices, music, and merriment (Cherry, 1953). To the more introverted reader, this is the same phenomenon as listening to a television show while the din of the air conditioning unit rattles on and the kid next door enthusiastically practices their new trumpet muffled only by the thin, shared apartment wall. While these situations of auditory streaming might seem so effortless as to be trivial, a look at the anatomy of sound and the auditory system will reveal the understated complexity of this ability.

A sound begins its life when an object vibrates, or repeatedly moves back and forth, causing movement in the surrounding air molecules which are displaced in alternating moments of pushing and pulling. This pattern of displacement creates repeating areas of higher and lower air pressure, called compression and rarefaction, which together form a sound wave (Figure 1.1).

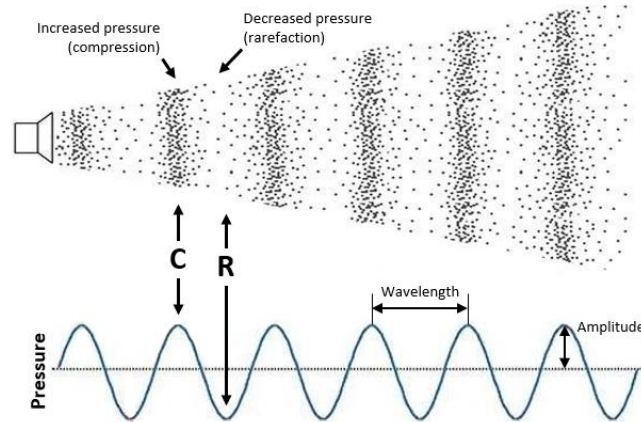


Figure 1.1: **Schematic diagram of a sound wave.** A speaker vibrates to produce fluctuating moments of high and low air pressure, compressions and rarefactions, graphically represented in the lower figure. Modified from (Deshpande et al., 2019).

You may notice that the depiction of compressed and rarefied air particles shown radiating away from the speaker doesn't look much like the classic wavy shape of a wave. This is because sound waves and what you're seeing is a longitudinal wave, or one where the air particles are being displaced in the same direction they are moving: away from the speaker (Berg, 2024). Most commonly (and intuitively), sound waves are graphically represented as the squiggling, up-and-down sinusoidal wave shown in the lower section of the figure, summarizing the fluctuations in air pressure moving away from the source. A look at the vertical, connecting arrows shows the relationship between the vertical position of the wave, or its phase, to the longitudinal wave above—the moments of compression correspond to wave peaks while the moments of rarefaction correspond to the lower wave troughs, with smooth transitions in between.

Visualizing a sound wave in this way helps us learn a bit about the anatomy of a sound and the implications its physical structure has on how it sounds. Wavelength describes the distance between two identical phases along the wave (in this case the distance between two peaks), which defines the pitch of the sound—the smaller the wavelength, the higher pitched the sound. Put simply, think about plucking a rubber band held tautly between your fingers which would result in a relatively high-pitched twang. This is because the band vibrates rapidly, completing a single vibration more quickly, and thus has a smaller

wavelength with a shorter distance between peaks. The same rubber band held slackly between your fingers, meanwhile, sounds low-pitched when plucked because the band, loosely flopping about, completes a single vibration slowly leading to a greater distance between peaks and a greater wavelength. With this in mind, low-pitched sounds are said to be low frequency because a longer wavelength means fewer vibrations take place over the same time period as a relative higher-pitched, or high frequency, sound.

Next, amplitude describes the wave vertically, in terms of its deviation from the average sound pressure—how peaky are the peaks. The variation in air pressure between peak and trough defines the amount of energy in a vibration which we perceive as the loudness of a sound: gently setting your coffee mug down on a coaster results in a quiet, low energy click, whereas dropping it results in a ruckus as the shattering parts vibrate with high energy as they tumble.

We can now think of the wave's movement through the air: it invisibly radiates from its source much like waves in water that radiate from where a pebble was dropped. When a sound wave moving through the air reaches the cartilaginous outer ear, called the pinna, of a listener it next enters the ear canal (Figure 1.2). At the end of the ear canal is the eardrum, or tympanic membrane, a thin tissue separating the outer ear from the middle ear. The eardrum moves in response to being struck by air pressure fluctuations of sound waves like, unsurprisingly, the head of a drum. In turn, vibration of the eardrum moves the three small bones of the middle ear, the ossicles, which then transfer the vibration to the fluid within the inner ear's cochlea, a snail-shaped cavity containing a membrane along its length. This membrane, the basilar membrane, is lined with sensory hair cells with bundles of slender tips that deflect—like seaweed in a gentle tide—as it and the surrounding fluid move in response to sound. The movement of these mechanically sensitive hair cell tips cause them to create electrical signals that are relayed by cells called neurons, the fundamental unit of the brain that communicate complex information using these small electrical signals. In this case, the brain will ultimately interpret this electrical activity of neurons as a sound.



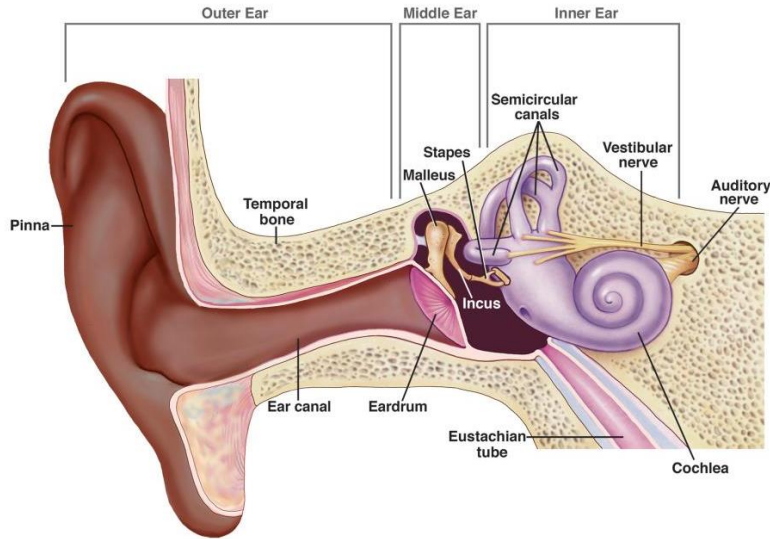


Figure 1.2: **Anatomy of the ear.** Diagram showing structures of the ear that facilitate the conversion of sound waves from air movement to electrical signals that travel to the brain. Here, the three bones comprising the auditory ossicles of the middle ear are labeled individually. Reproduced from niddc.nih.gov (Anon, 2022).

All sounds enter the brain via this same route—electrical signals generated by the physical movement of hair cells along the basilar membrane—which may well seem like an information bottleneck throttling complexity. However, there is an immense breadth and depth of sounds we can experience, from the thin simplicity of a coin sliding across the diner counter to the lush fullness of a symphony orchestra. To simplify this complexity to a single dimension like the differences in pitch we experience, a plucked guitar string sounds a lot higher in pitch than a plucked bass guitar string. One of the main reasons we perceive pitch because the hair cells along the basilar membrane are arranged tonotopically, or in a configuration where cells that preferentially are excited by high- or low-pitched sounds are laid out in a gradient (Figure 1.3) (Fettiplace, 2020). The basilar membrane is not uniform along its length. Instead, it is initially stiff at the base, tapering and thinning as it proceeds deeper into the coils of the snail shell. The stiff base of the cochlea is most sensitive to being displaced by rapid, high frequency vibrations while the flexible apex is most sensitive to slow, low frequency vibrations, with a smooth gradient in between. As a result, a high-pitched guitar note will initiate electrical activity in a different region of the cochlea than will a relatively lower note played on bass guitar. The brain then interprets this spatially organized electrical activity as pitch.

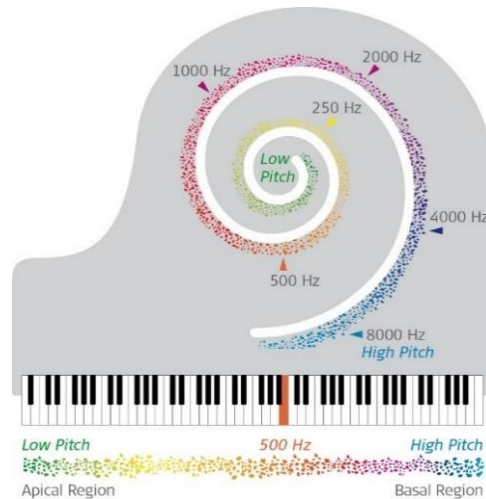


Figure 1.3: **Schematic of the tonotopic organization of the cochlear.** Simplified schematic depicting spiral geometry of the cochlea and the relationship between location along the length of the basilar membrane and pitch sensitivity. The basal region on the outside of the spiral is closest to where the ossicles first vibrate the cochlea and is most sensitive to high-pitched sounds, while the deep apical region of the cochlea is sensitive to low-pitched sounds. Reproduced from (MED-EL, 2017).

In a world where each individual sound waits its turn to vibrate the eardrum one at a time things would be very simple; each bout of electrical activity generated by the cochlea would be directly attributable to a single sound. Alas, here in reality any number of sound wave mixtures could strike the eardrum at any moment, necessitating auditory streaming. As a result, we arrive at the logistical challenge facing the brain that will also be at the heart of this dissertation: multiple sounds can simultaneously strike and vibrate the eardrum to generate infinitely complex patterns of electric signaling to the brain. So, how then do neurons of the brain tease apart and represent in their electrical activity the individual components of that noisy signal?

To complicate matters further, even individual sounds are typically far from simple. The simplest sound is the pure tone, a synthetic sound that cannot be generated by natural means. These tones result from a vibration at a single frequency and intensity (a pure sinusoidal waveform as in Figure 1.1) which we perceive as a drawn-out beep at a constant pitch akin to a dial tone. Everyday hearing is nowhere near this simple, however. With few exceptions, the sounds we routinely experience are natural sounds: the shrieking scrub jay in the tree outside, the urgent whistle of the passing freight train as it blocks your route to work again, or even the jackhammer breaking ground on Portland’s next opulent apartment complex. Natural sounds contain limitless variety imparted to them by the unique physical properties of the objects that generate them (Theunissen and Elie, 2014).

Each sound, whether a simple pure tone or a complex natural sound, can then be described by two dimensions. The first is temporal, simply describing how sound unfolds over time. The second is spectral, describing the high- or low-pitched content of a sound in terms of frequency. Remember, a sound starts its life as a vibrating object displacing the surrounding air in alternating moments of high and low air pressure. The speed or frequency at which the object vibrates determines how rapidly these alternating moments of pressure will occur and strike the ear drum, which in turn dictates where on the cochlea will the tonotopically arranged hair cells be deflected and therefore what we perceive as the pitch of a sound.

A useful way of visualizing these components of sound is using the spectrogram, a tool that illustrates which frequencies are present at a particular time in a sound, depicted in darker colors (Figure 1.4) (Lyon et al., 2010). In the upper row of the figure, spectrograms representing two-second-long excerpts of two different pure tones, low- and high-pitched, are shown. Remembering that pure tones have the persistent flatness of a dial tone, notice the connection between a sound and how it is visually reflected in a spectrogram: the horizontal black bands show that each tone’s respective pitch is a consistent drone as you read from left to right across time. Compare this now to the lower row of spectrograms, showing a sentence spoken by a woman with a Southern twang and a waterfall, both natural sounds. You can use your knowledge of what these might sound like to gain an intuition of how to “read” a spectrogram and to appreciate the complexity of natural sounds: unlike pure tones they constantly vary in time and frequency, with numerous frequencies occurring at any given moment in complex relationships to one another.

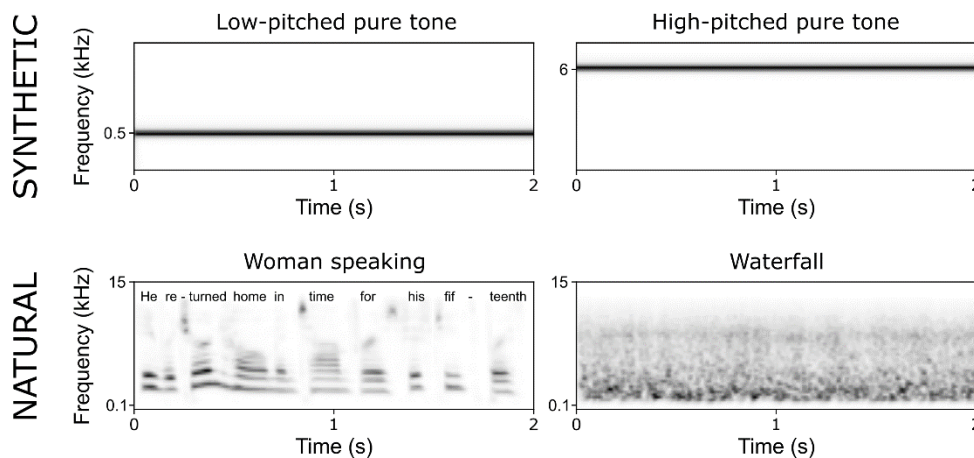


Figure 1.4: **Example spectrograms of pure tones and natural sounds.** Spectrograms are tools for visualizing a sound, showing how that sound unfolds from left to right across time as well as the lowness or highness of the pitch frequencies present in the sound (log-spaced bottom to top). Darker coloring indicates increased sound power at that frequency. The synthetic pure tones represented above are constant over time at a single pitch frequency. Natural sounds, by comparison, have much greater temporal dynamics and much more diverse frequency components. Imagine how the low, constant rumble of a waterfall produces

the spectrogram on the lower right, while the dynamic voice of a woman with a Southern drawl speaking has gaps in time that allow each word can be visually resolved.

To reframe this discussion around auditory streaming, look at the composite spectrogram of the same waterfall and woman talking simultaneously (Figure 1.5, *left*) and notice how the distinctive, clean features of each word of the sentence shown in Figure 1.4 are now less distinct, muddled by the roar of the waterfall. Now, a jaunty fiddle tune also begins in the background (*center*), further obscuring the speech spectrogram. Finally, add a jackhammer for good measure (*right*). At this point, it might only be possible to visually pick out words in the sentence because you remember what the clean speech spectrogram looked like previously, yet your brain might easily accomplish this feat with sound when performing auditory streaming. A recurring principle throughout this dissertation will be the exploration of how the components of these complex mixtures are reflected in the electrical activity of neurons in the brain. Put another way, can we find evidence in neural activity that the brain has managed to separate the woman's voice from the waterfall, or will it just be a jumble of signals created by the mixture of the two?

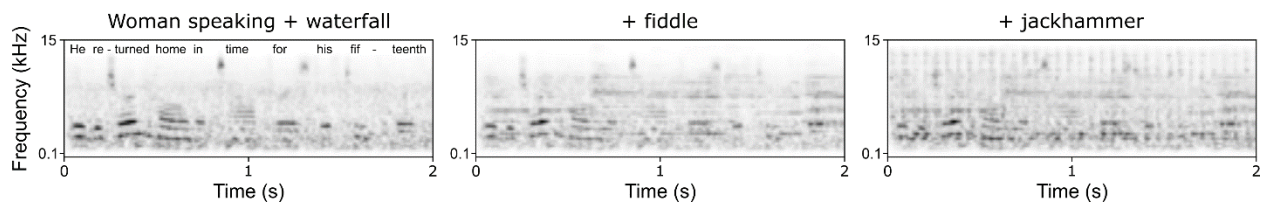


Figure 1.5: **Spectrograms of an increasing number of overlapping natural sounds.** At left, the spectrogram depicts the sounds of the waterfall and woman speaking from Figure 1.4 when occurring at the same time. Moving right, additional natural sounds of a fiddle and jackhammer are added into this mixture of natural sounds, visually illustrating the challenge of isolating a single sound source in auditory streaming.

Fortunately for us, despite this complexity the brain is particularly adept at making sense of sounds by relying on statistical regularities in the spectral and temporal dimensions that are unique to each sound (Moore and Gockel, 2012). Whether listening to a jazz trio or a symphony orchestra, you could easily follow a melody played by the flute because over time a flute always sounds like a flute due to its unique tone properties—its timbre—that sets it apart from other instruments. Our brains can track this spectral signature over time to group similar sounds as originating from the same source. Timbre is the one of the properties that helps you distinguish the voice of your conversation partner from all the other voices in a crowded room, revealing the importance of statistical regularities in the frequency domain for successful streaming.

Regularities in time also underlie successful streaming, where sounds that begin at the same time are more likely to originate from the same source. In a basic sense, a choir is impactful because many voices singing in unison creates the illusion of merging into a single vocal instrument where the individual voices cannot be segregated. As it relates to auditory streaming, the individual voices are unable to be perceptually separated due to strong cohesion in time, among other reasons. However, if the soprano voice errantly began singing half a beat off cue, the illusion of a unified vocal instrument is broken as she skews the timing to perceptually pop out of the stream. This shows us that our brains also use temporal cues to accomplish auditory streaming. These statistical regularities in the time and frequency domains that enable successful streaming are called grouping cues and interact in complex ways that will be discussed in much greater detail in forthcoming sections.

Now knowing some of the grouping cues your brain relies on when streaming, think back to the example in Figure 1.5, when the woman's voice was increasingly obscured by noise. It became quite challenging to isolate her voice even visually, yet your brain typically can identify even subtle spectral and temporal signatures of her voice effortlessly. Unfortunately, though auditory streaming is automatic for most listeners, hearing-impaired listeners often experience a greater challenge when it comes to hearing in noisy situations. In some forms of hearing impairment and even with certain hearing prosthetics, the electrical signal produced by the cochlea lacks the precise resolution necessary to preserve the subtleties of the grouping cues our brains rely on for accurate streaming.

For this reason, research that seeks to better understand how complex mixtures of natural sounds are represented in the electrical activity of neurons in the brain could give a clue as to the most important aspects of sounds that allow for effortless streaming. This knowledge could in turn guide the development of future hearing prosthetics to better preserve these essential features to improve auditory streaming in hearing impaired listeners (Bondy et al., 2004). The research to be presented in this dissertation aims to take early steps towards this long-term goal.

Hopefully the above introduction has provided a bird's-eye yet intuitive background of the peripheral auditory structures that allow us to interface with sound and how this leads to the immense challenge our brains face when making sense of complex natural sounds. Further, my hope is that this broad overview makes clear the motivation behind attempting to understand my central question of how neurons in the brain use electrical activity to represent complex mixtures of natural sounds in a way that allows us to make sense of our noisy world.

## **1.2 How does your brain make sense of sound in a busy acoustic world?**

In the previous section, we discussed that sound waves are the alternating compression and rarefaction of air resulting from the vibration of an object, movement that is translated by the machinery of the peripheral auditory system into an electrical signal ultimately routed to the brain. Even at the level of the periphery, it was possible to appreciate the complexity of auditory streaming particularly given the immense heterogeneity of natural sounds, where an entire soundscape must be reduced to electrical activity resulting from a complex pattern of cochlear activation. Throughout this next section, we will look in more detail at the journey sound takes once transduced to electrical activity by the cochlea as it travels through the brain.

To successfully stream and produce meaningful behavior, neurons need not only relay the received electrical signals, but they also need to disentangle mixes of sound and prune the signal by emphasizing features of relevant sounds to generate meaningful, noise-robust perception. Prior work has frequently investigated auditory streaming at the level of the computations performed by larger populations of neurons in the brain and at the level of perception using psychoacoustic studies. Here, we will explore the advantages and limitations of past approaches as part of a larger conversation contextualizing prior studies of auditory streaming as they relate to the work to be presented in this dissertation.

### **1.2.1 Anatomy and physiology of the central auditory system: from cochlea to cortex**

Before discussing the cortical processing of sound that gives rise to perception, it is important to first lay out the roadmap over which the sound-evoked electrical signals generated in the cochlea must travel. To fully appreciate this pathway and the computational challenges at the steps along the way, we will begin a more complete discussion of the auditory system at the cochlea and proceed through the central auditory system to the cortex (Figure 1.6).

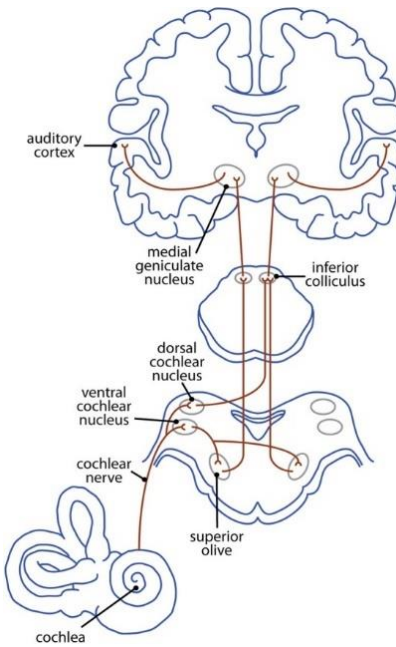


Figure 1.6: **The ascending auditory pathway.** Schematic of the ascending auditory pathway from cochlea to cortex. Reproduced from (Butler and Lomber, 2013).

### ***Cochlea***

As the tonotopically arranged hair cells along the basilar membrane of the cochlea (Figure 1.3) are deflected in response to vibrations, a sound is decomposed into its frequency constituents and transduced into a complex pattern of electrical signals (Fettiplace, 2020). This signal travels to the brain via the corresponding auditory nerve (AN) fibers, which preserve the tonotopy of the cochlea by encoding information from different sound frequency channels.

### ***Brainstem***

The complex electrical signal pattern carried by AN fibers form their first synapse in the ipsilateral (on the same side of the body as the referenced ear) cochlear nucleus (CN) of the brainstem. The ventral cochlear nucleus (VCN) projects both to the ipsilateral and contralateral superior olive. Here, our ability to localize sound in the horizontal plane of space emerges via the integration of binaural cues as the medial and lateral superior olive (MSO, LSO) receive and compare electrical signals from both ears.

By representing sound with high temporal precision, differences in the arrival time of sound at each ear, called the interaural time difference (ITD) can be computed by neurons of the superior olive (Stecker and Gallun, 2012). Because our ears are located on opposite sides of our head, a sound that occurs on our right side will arrive at the right ear slightly before it travels the width of the head to arrive at the left ear. The very slight (just a few 100  $\mu$ s) earlier arrival time in the MSO from the right ear cues the listener that

the sound is on their right. Similarly, a sound source directly in front of the listener is equidistant to both ears, and thus the ITD would be approximately 0  $\mu$ s.

Sound intensity, or loudness, is also a binaural cue arising in the superior olive. A sound coming from the right side of the listener not only arrives at the left ear later, but the head also casts a sound shadow that causes the intensity of the sound reaching the left ear to be slightly less than that of the right ear, a difference referred to as the interaural level difference (ILD) (Tollin, 2003). Classically, the ‘duplex theory’ has suggested that ITDs are facilitated by low-frequency sound localization in the medial superior olive, while ILDs calculated in the lateral superior olive (LSO) are used for high-frequency sound localization when frequency becomes too rapid to accurately detect slight phase shifts (Bernstein and Trahiotis, 1985). More recent evidence has suggested a role of the LSO in ITDs (Franken et al., 2021). Together, the superior olive of the brainstem is critical for the processing of binaural cues that aid in horizontal-plane sound localization.

### ***Midbrain/subcortical***

Activity from the ipsilateral dorsal cochlear nucleus (DCN) and contralateral superior olive next synapse in the midbrain at the inferior colliculus (IC). In addition to playing a role in the ascending and descending auditory pathway, the IC is notable due to its prominent role in multisensory integration and context-related connections, with IC cells having been reported to be sensitive to visual, oculomotor, somatosensory, behavior, and reward signals (Gruters and Groh, 2012).

Neurons in IC also have firing properties that allow for rapid, low latency synaptic transmission that can lock to the onsets of sound, creating a spike timing code (Trussell, 1999). This code is found at earlier stages in the auditory pathway and allows for the encoding of precise neural timing of stimuli through onset first spike latency which can aid in computations such as sound source localization from binaural cues (Furukawa et al., 2000). As temporal integration windows increase across the auditory hierarchy limiting the ability to lock timing as precisely, the spike timing code is largely supplanted by a rate code to adjust firing rate to reflect amplitude modulation and change in the stimulus (Lu et al., 2001; Niwa et al., 2012).

Auditory information with precise timing from the IC continues its ascent of the auditory pathway to synapse in the final subcortical region, the medial geniculate body (MGB) of the thalamus. While the role of the MGB has traditionally been thought of as a relay center that simply projects to the auditory cortex (AC), the MGB is now recognized to have a more nuanced role whereby it filters incoming inputs to modify the representation of acoustic features to be used by the AC. MGB neurons have been implicated in rapid feedforward inhibition, facilitation of excitatory inputs, feature extraction, feature integration, feature learning, and integration of sensory information (Bartlett, 2013).



## Cortex

The MGB projects extensively to the auditory cortex (AC), the primary destination of auditory information in the ascending pathway. The AC is subdivided into the primary AC (A1) and peripheral, or belt, areas (Purves et al., 2001). A1 is located on the superior temporal gyrus of the temporal lobe and receives extensive projections from the ventral thalamus. The tonotopy of the cochlea is preserved in the thalamus and thalamocortical projections to A1, which similarly is organized tonotopically (Figure 1.7) (Zhang et al., 2001). Meanwhile, belt areas of AC lack the point-to-point inputs from the thalamus that preserve tonotopy, leading to looser tonotopic organization.

The cortex is divided into laminar layers based on properties such as cell types, outputs, and inputs. The primary inputs from the ventral thalamus terminate mainly in cortical layer 3/4 (Lee, 2013). Inputs to A1 can either be transmitted to the belt regions or connect back to different layers of A1 (Hackett, 2011). While primary and belt regions of the AC can be defined hierarchically based on this projection architecture, they can also be defined by their response properties, where belt regions show increased response latency and broader frequency tuning (Bizley et al., 2005; Atiani et al., 2014; Norman-Haignere and McDermott, 2018). Further, neurons in the AC can be tuned to respond to increasingly complex and abstract statistical spectrotemporal relationships (Rauschecker et al., 1995; Kikuchi et al., 2014), a topic which will be extensively discussed in Section 1.2.5.

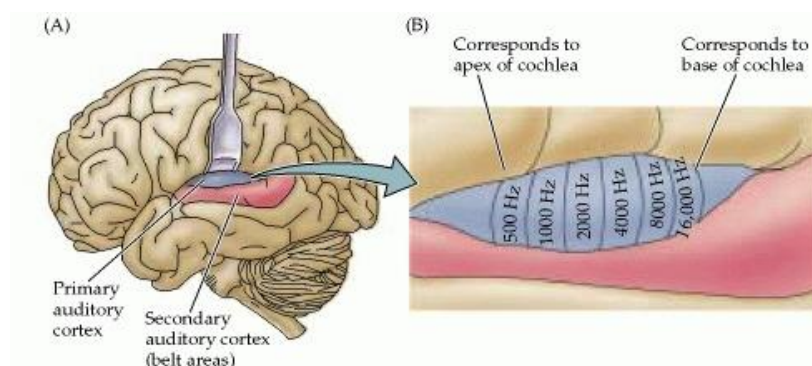


Figure 1.7: **The human auditory cortex (AC).** *A*, Diagram of the location of the primary (A1, blue) and belt (red) areas of AC on the superior temporal gyrus. *B*, Details of the preserved tonotopic organization of A1, which is preserved and inherited from the cochlea. Reproduced from (Purves et al., 2001).

## 1.2.2 Types of sound stimuli

Having outlined the extensive steps over which auditory information travels en route to the AC, it becomes possible to appreciate the distance and extent of the transformations these signals must undergo

before the cortex interprets the signal. Further, in the context of auditory streaming, information on both relevant and irrelevant sounds will largely be preserved throughout the ascending auditory pathway. Before discussing how the AC represents the incoming information and permits behavior in the context of streaming, it is helpful to review some properties of synthetic and natural stimuli, the two flavors of sound stimuli used throughout the auditory streaming literature.

As discussed in Section 1.1, heterogeneity is characteristic of natural sounds, allowing listeners to distinguish a limitless number of unique sound sources whether they vary subtly or hugely in spectral and temporal properties (Figure 1.4). More precisely, natural sounds possess a power law relationship that confers spectro-temporal correlations over multiple timescales. This relationship distinguishes them from synthetic sounds, which can range from being completely uncorrelated as in white-noise signals, or singly correlated as in pure tones generated by a sinusoidal wave (Theunissen and Elie, 2014). Compared to synthetic sounds, natural sounds are information-rich and perceptually sophisticated, permitting complex behavioral responses and neural coding.

As a result of the immense spectral and temporal variety in natural sounds, streaming cannot be accomplished simply by differentiation based on the tonotopic channels inherited from the cochlea. Instead, streaming of natural sounds requires distinguishing stimuli according to grouping cues, statistical regularities in the time and frequency domains (Bregman, 1990).

While synthetic sounds can be as simplistic as the pure tone or other uncorrelated noise signals, more recent sound synthesis methods allow artificial sounds to be synthesized to possess statistical properties that mimic naturalistic cochlear activation patterns and response properties (Popham et al., 2018; Shearer et al., 2018). The extent to which synthetic sounds can be curated to contain natural properties can extend anywhere from engaging similar tonotopic activation on the cochlea to modeling natural spectrotemporal relationships (Norman-Haignere and McDermott, 2018). Throughout the literature as well as in the forthcoming discussion of results in Chapter 2, synthetic sounds have an incredible breadth of applications afforded to them by this large flexibility of properties. Still, we will also see that no matter how closely and to what extent the complexity of synthetic sounds approximate a natural sound, they will always lack the fully natural statistics that give natural sounds that *je ne sais quoi* our auditory systems are particularly well-equipped to discriminate (Młynarski and McDermott, 2019).

### **1.2.3 Population level and psychoacoustic studies of auditory stream formation in humans**

Auditory streaming is a phenomenon particularly amenable to psychoacoustic studies, which bring together the physiology and the psychology of sound to systematically probe principles of auditory stream formation with the direct readout of human perception and behavior. Classically, a reductionist approach has allowed probing of the fundamental properties of auditory streaming by favoring simple, synthetic sound stimuli with tractable properties that can be isolated, controlled, and replicated (Bregman et al., 2000; Shearer et al., 2018). Using these stimuli, psychoacoustic experiments have highlighted the importance of and sought the perceptual boundaries of spectral (Cusack and Roberts, 2000; Popham et al., 2018; McPherson et al., 2022), temporal (Andreou et al., 2011; Shamma et al., 2011; Sollini et al., 2022), and spatial (Akeroyd et al., 2005; Middlebrooks and Onsan, 2012; Bizley and Cohen, 2013) sound statistics to successful streaming. Alternatively, more ethologically oriented approaches have favored using natural sounds to examine the behavioral outputs of streaming while the auditory system is faced with more natural acoustic contexts (Norman-Haignere and McDermott, 2018; Młynarski and McDermott, 2019).

In the following review-style section, we will explore how psychoacoustic and population-level studies in humans have informed descriptions of auditory stream formation, with the goal of highlighting foundational grouping principles that will be relevant throughout the dissertation. I will also weigh the relative benefits and limitations of synthetic and natural sound stimuli. By gaining a background of how auditory streaming has been studied in humans and the kinds of questions that can be answered with these approaches, I will contextualize the motivation behind my own experimental choices, as they relate to the questions posed in this dissertation.

#### ***Synthetic sound stimuli***

No discussion of auditory streaming would be complete without the classic synthetic stimulus configuration for auditory stream formation, the ABA– paradigm (Bregman, 1994). This stimulus structure explores relative sound properties in the spectral and temporal domains that lead to a sequence of pure tone stimuli to be heard as a single, coherent sound stream or multiple segregated streams. Here, both A and B are brief pure tones followed by a silent period (‘–’) of equal duration. As the frequency interval separating A and B ( $\Delta F$ ) diminishes or as the time rate at which the notes are presented ( $\Delta T$ ) grows, the fusion of streams A and B into a single, galloping rhythm is perceived. Alternatively, expanding  $\Delta F$  or shrinking  $\Delta T$  results in the percept of two separate streams, with stream A proceeding at twice the rate of stream B (Figure 1.7). This finding uses the most simplistic pure tone stimuli to establish

spectro-temporal relationships ( $\Delta F$ ,  $\Delta T$ ) that underlie the percept of stream fusion and segregation in a descriptive manner.

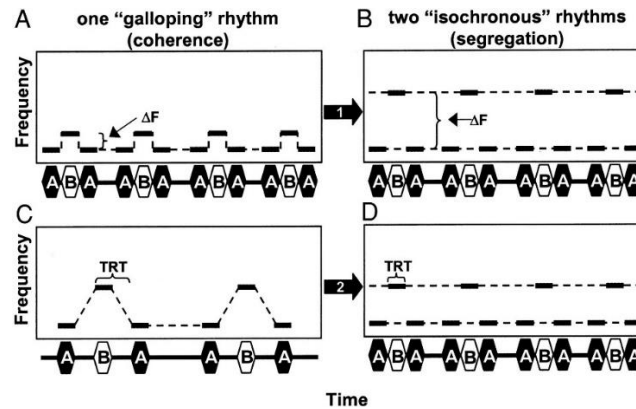


Figure 1.7: **Schematic of the ABA-paradigm.** The top row illustrates the relationship of increasing frequency separation between A and B ( $\Delta F$ ) in fusion (here called “coherence”) and segregation while the bottom row shows the relationship of decreasing the temporal separation between A and B ( $\Delta T$ ). Dotted lines connecting tones on the spectrograms indicate the perception of one auditory stream (coherence) or of two streams (segregation). Reproduced from (Bee and Klump, 2004).

Illusions can also be a useful tool in sensory neuroscience, allowing guided curation and presentation of synthetic sound relationships to force a perspective by leveraging known principles. The ‘octave illusion,’ first described by Diana Deutsch, refers to sound stimuli consisting of two pure tones, separated by an octave, presented alternately but in opposite phase between the ears. Typically, this yields the percept of all the low tones lateralized in one ear alternating with higher tones in the opposite ear, both at half the presentation rate (Figure 1.8a) (Deutsch, 1974). While a mechanism separating pitch determination and sound localization was proposed at that time, a more recent study aimed to refine and challenge this understanding of the illusion, framing it instead as an auditory streaming problem (Mehta et al., 2017). Noting that the octave illusion shares features such as a minimum  $\Delta F$  (as in the ABA-paradigm) and having attention- or instruction-based effects, the goal was to empirically resolve understandings of the illusion.

A psychoacoustic study was used to probe the neural correlates of the octave illusion, adding a brief series of monaural precursor cues prior to the classic stimulus as well as amplitude modulation or fading to the uncued tone in the contralateral or ipsilateral stream relative to the cue (Figure 1.8b,c). Placing these modulations synchronously (contralateral and concurrent) to the cued tone revealed the illusion was perceived with these modifications. These results implicated a mechanism whereby the illusion is formed due to a misattribution of the timing of synchronous tones, rather than the prior assumption that the

temporally alternating tones were being spatially misattributed, thus revealing a role of temporal synchrony in auditory stream formation.

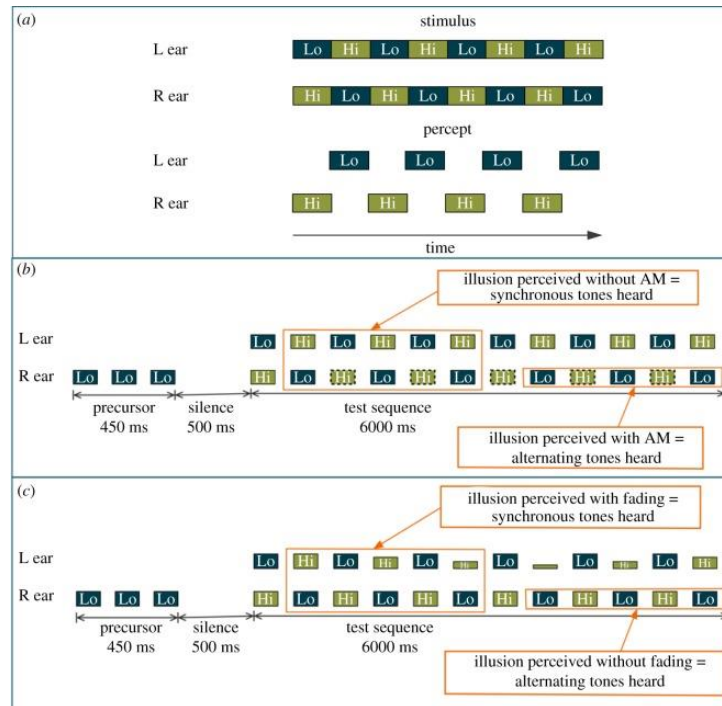


Figure 1.8: **The octave illusion.** (A) Schematic of the original octave illusion as described by Diana Deutsch in 1974, showing both the stimulus presented and the most common perception. ‘Lo’ refers to the low-frequency tone and ‘Hi’ to the high-frequency tone. (B-C) Schematics of the variations of the octave illusion described in the main text. Reproduced from (Mehta et al., 2017).

So far, we’ve discussed stimuli composed of pure tones arranged in a variety of clever spectral, spatial, and temporal configurations, allowing basic principles of auditory stream formation to be leveraged with this simple, flexible tool. Still, a major limitation of synthetic sound stimuli is limited ethological relevance. In particular, pure tones differ from more spectrally complex natural sounds, or even synthetic noise stimuli, in that they activate a discrete, narrow region of the cochlea, removing any contextual response interactions when more broad or correlated regions of the cochlea are sampled. As such, we will shift our focus to more statistically complex synthetic sounds that retain the benefits of a curated, consistently replicable synthetic stimuli.

Iterated ripple noises (IRNs) were used to explore spectral correlations as they relate to auditory stream formation in normal-hearing and hearing-impaired listeners, a group particularly challenged in acoustically cluttered environments (Shearer et al., 2018). In brief, an IRN is a noisy signal that is given perceived spectral content through the addition of a signal filtered at different frequencies. This results in

a signal that contains both noise and repetitions of spectral information in different time delays, hence “ripples,” which can be perceived as a frequency based on the relative periodicity of the temporal waveform. The extent of the ripple delay confers the perception of pitch and the number of iterations of added filtered noise correlates with increased perception of pitch strength. The spectral content of IRNs is like harmonic complexes – a synthetic, periodic sound stimuli composed of spectral stacks of pure tones at harmonic intervals that also contain perceived pitch identity.

Here, the ABA– paradigm was replicated using IRNs with the expectation that stimuli with greater pitch strength (more iterations, less high-pass filtering to conserve spectral information) will be increasingly likely to segregate into two streams by both groups, with hearing-impaired listeners segregating the alternated IRNs less readily as a result due to deficits in the processing of temporal and spectral cues that confer pitch to IRNs. Indeed, results were consistent with this hypothesis and show that auditory stream segregation is contributed to by factors such as pitch difference ( $\Delta F$ ), spectral resolution (stimulus filter), temporal pitch (delay), and tone strength (number of iterations), the perception of which are all diminished in hearing-impaired listeners. A key aspect of the ethological relevance of IRN stimuli is they pair a tone with the kind of broadband cochlear activation one would expect in a more naturalistic sound. The brain must perform more complex computations to extract and isolate salient pitch information than simply relying on discrete tonotopic channel activation, thus more closely replicating a naturalistic auditory streaming representation.

The adaptation of the ABA– paradigm to further illuminate principles of auditory streaming can serve as an illustration of two points: 1) synthetic stimuli are capable of producing a diverse range of stimuli which psychoacoustic studies can use to probe different aspects of stream segregation and 2) by increasing the naturalistic quality of synthetic stimuli through the introduction of more complex statistical relationships, an increasingly nuanced view of auditory streaming can be obtained. Together, these highlight both a major advantage and limitation of using synthetic stimuli to study auditory streaming.

### ***Natural sound stimuli***

In comparison to synthetic sounds, natural sound stimuli carry the benefits of greater ethological relevance and complex spectral relationships varying constantly in time, resulting in naturalistic excitation along the auditory pathway. Speech is a dynamic sound loaded with behaviorally relevant information and is therefore an excellent place to begin a discussion about natural sounds and auditory streaming. Citing a dearth of studies that explicitly demonstrate the transfer of streaming grouping cues to ethological applications, speech is used to probe the acoustic grouping cue harmonicity (Popham et al., 2018). Harmonicity refers to the quality of sound when frequencies are integer multiples (harmonics) of a

common fundamental frequency (F0) and is also a crucial component of pitch perception, sound identity (timbre), and musical harmony (Theunissen and Elie, 2014).

Here, single-talker speech was used to study, within a single voice, spectral components that enable a statistically complex signal to fuse and be tracked over time. To do this, the third harmonic (F3) of a natural speech utterance was mistuned (Fig. 1.9a, b) and listeners were instructed to report how many sound sources they heard. The threshold for the mistuned harmonic segregating from the percept of the speech stimuli was just above a 2% mistuning of F3 (Figure 1.9c). Having determined harmonicity's contribution to grouping in individual speech utterances, this paradigm was adapted to a classic multi-talker psychoacoustic study to determine how inharmonic speech affects intelligibility in a multi-talker scenario. Intelligibility, measured by sentence comprehension, was decreased with progressive inharmonicity, demonstrating a role for classic acoustic grouping cues in an ethologically relevant scenario.

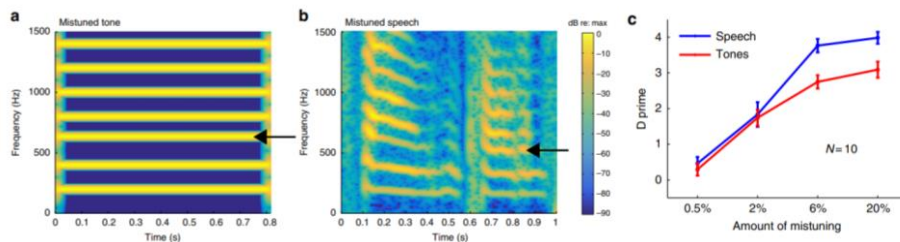


Figure 1.9: **Mistuned harmonics reveal the role of harmonicity in auditory stream formation.** (a-b) Example spectrograms showing a harmonic tone complex and natural speech with an upwardly mistuned third harmonic by 6%. (c) Sensitivity to mistuned harmonics in tones and speech. Very slightly mistuned harmonics continue to yield the perception of a unified tone or speech bout, while increasing the mistuning above 2% begins to increase the likelihood of a listener reporting the mistuned harmonic as segregating from the tone or speech. Reproduced from (Popham et al., 2018).

The previous study used natural sound stimuli to explore principles gleaned from psychoacoustic studies by generating precisely curated, synthetic versions of natural stimuli. Natural stimuli also have relevance because of the limitless range of statistical diversity the brain must generalize to invariantly select behaviorally relevant stimuli. To this end, pairing psychoacoustics and *in vivo* electrocorticography allows mechanisms of speech streaming in the presence of changing background noise to be explored (Khalighinejad et al., 2019). Here, patients undergoing chronic intracranial encephalography (iEEG) listened to continuous speech while competing natural background noises periodically changed was played to. Three background sounds with unique spectro-temporal properties were chosen. The result was electrophysiological responses that led to reconstructed spectrograms resembling the background

immediately following a noise transition, with a quick adaptation to return to a noise-invariant representation of the concurrent speech utterances. To leverage this result into a psychoacoustic task, listeners were to discern isolated phonemes as background sounds changed, showing decreased phonetic identification during the adaptation period relative to the period after adaptation (Figure 1.10). These consistent neural and perceptual effects suggest a mechanism by which the statistics of natural distractors can be quickly adapted away to permit robust, noise-invariant perception of behavioral salient stimuli.

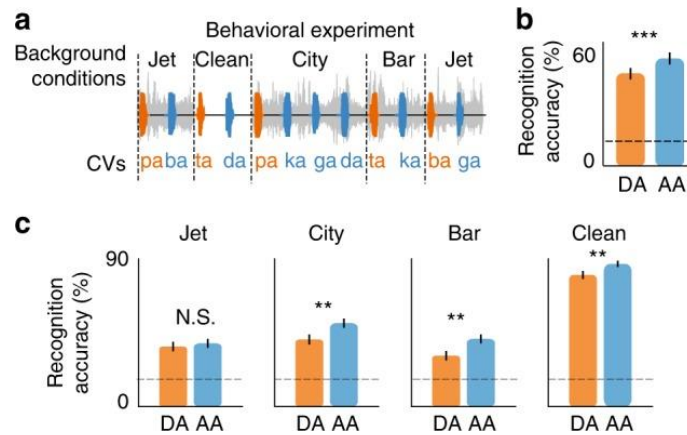


Figure 1.10: **Statistically unique natural background noise differentially affects phoneme recognition.** (a) Diagram of task structure whereby a listener must identify phonemes in differing natural background noise conditions. (b) Recognition performance is substantially increased following a period of adaptation (after adaptation, AA) relative to the period immediately following a background transition (during adaptation, DA). (c) Results from *b* broken down by individual background condition. Statistically distinct background noise conditions obscure phoneme recognition to different extents, revealing the role of unique natural sound statistics in auditory streaming. Reproduced from (Khalighinejad et al., 2019).

The ethological relevance of utilizing stimuli like speech and environmental sound textures that carry intrinsic behavioral salience and daily relevance is clear and reveals a major benefit to using natural sounds when searching for a mechanistic explanation of a perceptual phenomenon. Moreover, the differential magnitudes of phoneme obscuration during the adaptation phase for different backgrounds highlights the importance of the unique spectro-temporal statistics inherent to each natural sound, which will be a major focus of the forthcoming results presented in Chapter 2 of this dissertation.

Thinking about natural statistics raises a challenge though, as a limitless number of unique spectro-temporal statistical patterns are present and constantly varying in natural sounds and therefore all configurations cannot be controlled or exhaustively be presented to an experimental subject. Here, we arrive at a limitation of using natural sounds to study auditory streaming; the need to generalize stimuli, the forte of synthetic stimuli. Though I have thus far presented synthetic or natural stimuli as though they



were a binary choice where generalizability comes at the expense of ethological relevance, efforts have been made to bridge this gap and allow the advantages of synthetic and natural stimuli to complement one another, maximizing the respective explanatory power of each.

### ***Synthetic and natural sound stimuli as complementary approaches***

A study demonstrates a holistic approach to stimuli selection of natural sounds, reducing their staggering complexity in a unique and clever way (Młynarski and McDermott, 2019). The study used convolutional sparse coding on a corpus of speech or solo musical instrument (the model relies upon a huge set of examples) cochleagrams to learn a set of features, or spectrotemporal motifs, that appear within natural sounds. Such features include simple and spectrotemporally local patterns—single frequencies, harmonic frequency sets, clicks, and noise bursts—that the early auditory system responds to (Figure 1.11A). The challenge in auditory streaming is grouping all these smaller features within a sound source to provide continuous perception. As such, the derived features each had their relative co-occurrences calculated, quantifying how likely each feature appears in relation to others within a natural sound. The hypothesis was that more co-occurring features share properties that improve the likelihood of being perceived as originating from a single source due to listener internalized statistical relationships.

To test the hypothesis that listeners have an internalized model of these co-occurrence statistics, sound features with high or low co-occurrence likelihoods were presented concurrently and listeners reported whether a trial contained a single or two sound sources. Features likely to co-occur were reliably identified by listeners as from a single sound source, while those less likely to co-occur were often perceived as two separate sounds (Figure 1.11B). Importantly, a control in which synthetic sound textures were used to generate the feature library led to near chance performance, indicating natural sounds are critical to the internalized statistical model used by listeners (Figure 1.11C). These data provide a case for a viable reductionist approach to natural sounds, while further emphasizing the importance of natural statistical relationships uniquely present in natural sounds.

Further, grouping cues were derived from the co-occurrence statistics and reiterated several known grouping cues where features with similar spectral content or with common temporal onsets/offsets of  $F_0$  are likely to co-occur. These can be seen in our discussion of the properties of  $\Delta T$  and  $\Delta F$  that lead to such percepts as the ABA– paradigm, temporally synchronous tones generating the octave illusion, or the inharmonic speech and tone complexes. Another derived grouping cue however, spectrotemporal modulation—features with difference spectral shapes (tones versus clicks) tend to not co-occur—had gone previously unreported in the auditory streaming literature likely due to its less intuitive nature. Still, this cue proved as perceptually salient a predictor of stream formation as derived cues that reiterate known cues, underscoring the importance of ethologically relevant sound selection when approaching an

understanding of an ethologically relevant sensory task such as streaming by highlighting the breadth of complex statistics within natural sounds as well as listeners’ internalized model of natural sounds, a theme that will recur throughout Chapters 2 and 3 of this dissertation.

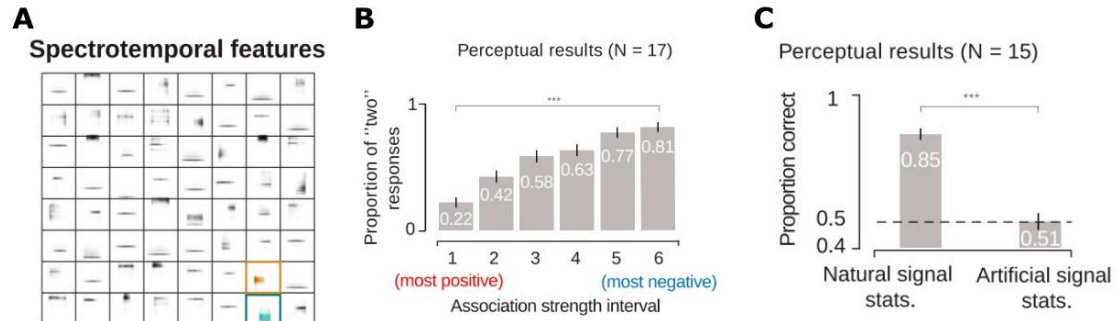


Figure 1.11: **Naturally co-occurring spectrotemporal features are more likely to be perceptually grouped.** **A**, Examples of model-learned spectrotemporal features. **B**, Results from a psychoacoustic experiment where spectrotemporal features were paired and listeners reported whether they heard a single or two separate sounds. ‘Most positive’ refers to spectrotemporal features with high co-occurrence statistics, while ‘most negative’ refers to features with low co-occurrence statistics. **C**, Results from a psychoacoustic experiment where co-occurring features learned from natural sounds or synthetic sounds were paired. Artificial features produced chance performance, spectrotemporal grouping judgements are specific to natural sounds. Reproduced from components of (Młynarski and McDermott, 2019).

Despite this, a reductionist approach and synthetic stimuli are not without value. We have already looked at the details as well as an overview of mechanistic explanations that have been identified using simple, artificial stimuli. We’ve also discussed a limitation of natural sounds by which an endless library of spectrotemporal possibilities makes sampling that space an experimentally difficult task that, as we just saw above, requires reduction of its own to become broadly informative. Thus, it is important to understand that the information required for natural neural representation of acoustic signals is lost as the complex statistical relationships inherent to natural sounds are reduced.

A demonstration of this principle takes place outside of the auditory streaming literature, comparing fMRI responses to natural sounds and to model-matched synthetic stimuli generated by conserving different statistics of the natural stimuli (Norman-Haignere and McDermott, 2018). Responses in primary regions of auditory cortex were comparable between the natural and synthetic versions fully model-matched to natural spectrotemporal relationships. However, responses to these synthetics in non-primary regions diverged, suggesting an auditory hierarchy in which secondary regions respond to higher-order statistics not able to be captured by the model (Figure 1.12). To take this further, synthetics can also be selectively generated using either the filter that matches the spectral or temporal domains, or eliminating

both leaving a synthetic sound that only matches the cochlear activation pattern of the original sound. With this increasing randomization of natural statistics, responses of primary regions to the synthetics no longer match the original. This change demonstrates that even at the level of primary auditory cortex natural spectral and temporal correlations are key to driving natural responses, a point that will be reflected in a key analysis in Chapter 2.

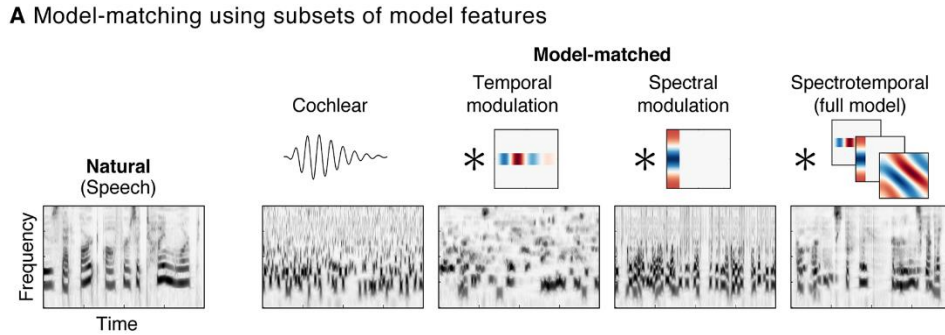


Figure 1.12: **Natural stimuli with model-matched synthetic stimuli.** An example natural sound cochleagram shown with various model-matched synthetic stimuli, moving from the most degraded to a full model preserving modulation statistics of the original. Reproduced from components of (Norman-Haignere and McDermott, 2018).

### **Takeaways**

Taking this together, it is crucial to tailor the sound stimuli being used to the scope of the question being asked. When asking questions about the perception of auditory stream formation, it is very reasonable to use psychoacoustic studies where stimuli can be manipulated. Similarly, when attempting to uncover the effect of complex sound relationships on stream formation, using natural sounds incrementally modified to alter perception makes sense. Still, the discussion so far has largely centered around psychoacoustic studies or population electrophysiology using non-invasive methodologies. For a sensory processing task like auditory streaming, this has allowed a huge range of information to be gained descriptively and mechanistically. At the same time, due to the relative inaccessibility of large-scale, invasive electrophysiology techniques in humans we can only uncover mechanisms up to a limited resolution, preventing studies of mechanisms of auditory streaming at the single-unit, representational level, a key question of this dissertation. Therefore, we will next explore what can be gained by trading away the behavioral output of perception permitted in human studies, transitioning to studies of auditory streaming in animal models.

### **1.2.4 The role of animal models in studies of auditory streaming**

We have now had a sampling of the auditory streaming literature in human subjects. While these studies gave anywhere from a descriptive overview of the formation of the phenomenon to a more nuanced view of brain activity as it perceptually occurs, access to information about streaming at the single-unit resolution has been notably absent. Details on how and if specific sound features comprising mixed sounds are represented in individual neurons could reveal further mechanisms by which noise-invariant perception arises between the cochlea and perception. To answer these questions, animal models amenable to invasive electrophysiological recordings are necessary.

Animal studies of auditory streaming will not be without their drawbacks; we of course cannot directly ask an animal if they hear one or two streams in response to our manipulations as we can a human subject. There is, however, sufficient evidence of streaming-like behaviors in a number of species established as animal models—fish (Fay, 1998), birds (Hulse et al., 1997; Narayan et al., 2007; Dent et al., 2016), rodents (Noda and Takahashi, 2019), primates (Fishman et al., 2001; Izumi, 2002; Christison-Lagay and Cohen, 2014)—that would make these models a valuable tool to investigating auditory streaming at the level of single cell representation (Itatani and Klump, 2017).

In another review-style section, we will first survey studies of auditory streaming in animal models with the goal of highlighting unique methodologies and, once again, the effect of synthetic background noise versus noise with natural statistics. We will assess the strengths and drawbacks of each with respect to our discussion of the human auditory streaming literature and, ultimately, use this discussion to guide the methods we chose our own studies. Following this more sweeping review of the literature, we will conclude with a brief background and specific rationale for the animal models I chose for the work presented in this dissertation.

#### ***How has auditory streaming been studied in animal models using synthetic stimuli?***

We will open the discussion with two studies using ferrets which probe the robustness of neural responses to natural speech and ferret vocalizations when masked by synthetic background noise. Beginning with the perceptual experience of auditory streaming—that behaviorally relevant sounds can be identified in background noise—the first study aimed to identify where and how noise-invariant representation of relevant sounds arises in the auditory system (Rabinowitz et al., 2013). Citing evidence that background noise impacts behaviorally relevant sounds by altering their statistics, the authors put forth a hypothesis whereby noise-invariance could result from neurons along the auditory hierarchy (Figure 1.6) increasingly adapting to noisy sound statistics to progressively filter it out.

The study recorded neural responses in the inferior colliculus (IC) and AC of anesthetized ferrets while natural speech was played over stationary synthetic noise at varying relative sound levels. Statistically homogenous noise was chosen to make the distinction between background noise and the more behaviorally salient natural sound as unambiguous as possible. As expected, neural responses were progressively more invariant to noise level relative to the signal along the auditory pathway from IC to AC (Figure 1.13).

Noise can affect neural responses to a signal by increasing the baseline intensity, obscuring the signal by lowering the contrast between useful spectrotemporal information and the elevated baseline. Adaptation to the statistics of noisy stimuli would arise from a renormalization of responses relative to the noisy baseline, with neurons in areas further along the auditory hierarchy displaying more robust adaptation to intensity and contrast. Indeed, the introduction of noise in this study progressively caused less change in the firing rate along the auditory hierarchy, indicating a growing adaptation to changes in increased baseline intensity and decreased contrast, consistent with normalization that preserves relative contrast. These results provide evidence for increasing noise invariance and implies the emergence of auditory stream formation across the auditory hierarchy at the single-unit level as noise is progressively filtered out, implicating a mechanism of increased adaptation to the stimulus statistics of synthetic noise.

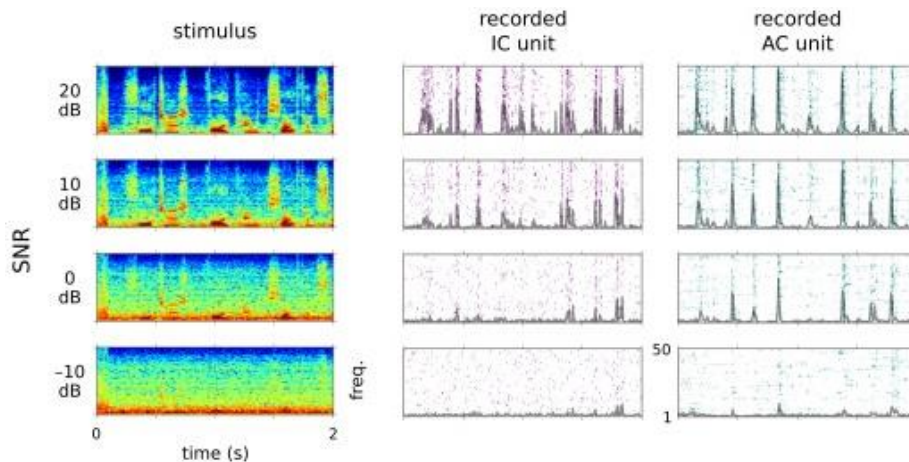


Figure 1.13: **Single unit responses to clean and noisy speech in inferior colliculus and auditory cortex.** Example single-unit responses in IC (center) and AC (right) to increasingly noisy speech signals (left). AC shows firing patterns more similar to the clean signal indicating a greater robustness to noise at this higher level of the AC. Modified from (Rabinowitz et al., 2013).

To further characterize neural mechanisms underlying noise-invariant signal representations in AC, another study recorded AC units from awake, passively listening ferrets during presentations of speech or ferret vocalizations with various synthetic background noises (Mesgarani et al., 2014). The study used synthetic maskers with uncomplex spectral profiles and stationary characteristics to simplify interactions with a diversity of speech foregrounds, these being white and pink noise. White noise contains an equal balance of power across all frequency channels like radio static. Pink noise similarly contains power across all frequency channels but the relative power decreases incrementally with higher frequency, resulting in a low rumbling like a waterfall. Reverberation, a temporal smearing of the original stimulus, was also included and has the same effect as noise of masking precise spectrotemporal characteristics.

To directly examine what aspects of the noisy stimulus are preserved in noise and to further bolster the observation of noise-invariant AC responses, the study used population reconstruction techniques. These techniques use the responses of populations of neurons to reconstruct a stimulus spectrogram that can be directly compared with the original. Consistent with the theory noise-invariant representations of the natural foreground sound, reconstructed spectrograms reliably resembled the original clean spectrograms, even in the presence of assorted synthetic background noise (Figure 1.14).

To explore neural mechanisms of this noise-robust representation in A1, receptive fields measured from the neural data were used to simulate neural responses. The receptive field serves as a filter that describes the transformation between stimulus and response. Additional transformations can be added to model additional dynamic biological processes. As such, a self-normalization mechanism modeled by combining feedforward synaptic depression (SD) and feedback gain normalization (GN) was able to accurately predict distortion effects of both the noisy and reverberant conditions. In contrast, a static model lacking these dynamic biological transformations, or either SD or GN alone, was insufficient to fully predict distortion effects, suggesting both are necessary for a mechanism of noise-robust signal representation.

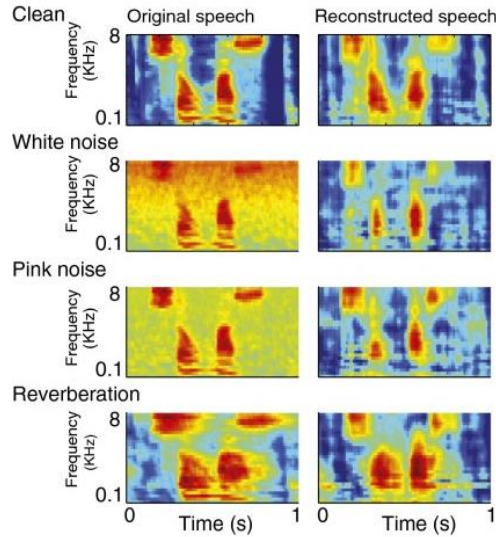


Figure 1.14: **Original and reconstructed spectrograms of clean and noisy speech.** (*Left*) Spectrograms of the clean or noisy speech played to the ferrets. (*Right*) Spectrograms reconstructed from population responses in A1. In all cases, the reconstructed spectrogram more strongly resembles the original clean speech than the features of the noise. Reproduced from (Mesgarani et al., 2014).

Taken together, these studies complement one another; both use similar methodologies of using a contrast between synthetic background noise and a natural foreground sound to study mechanisms by which noise-invariance arises in neural representations in the auditory pathway. The mechanisms proposed also both implicate self-normalizing adaptation processes that strip away the more regular statistics of competing noise. They also suggest the emergence of noise-invariance in single-unit representations whereby behaviorally relevant sounds are robustly and preferentially encoded in the presence of noise, the encoding of which typically fell by the wayside—just because a clean speech spectrogram can be recovered from population responses, does not necessarily mean information about the background was not also encoded. These studies also notably used synthetic background noise which, even when synthesized to be more naturalistic, result in dissimilar encoding (Norman-Haignere and McDermott, 2018). As a result, we will next look at a more limited body of single-unit studies of streaming that model streaming using natural foregrounds and backgrounds.

### ***How has auditory streaming been studied in animal models using natural stimuli?***

Using a natural foreground/background contrast, several studies have found results that challenge the dominance of noise-invariance at the single-unit level, suggesting that information about the noisy sounds is robustly encoded in the auditory system. A first study recorded single-unit activity in the zebra finch homolog of A1 during the presentation of a target birdsong with and without assorted background noise (Narayan et al., 2007).

Here, a conspecific birdsong was used as a target to be embedded in noise. Noise belonged to three categories: 1) broadband noise, a synthetic sound, 2) modulated noise, broadband noise modulated by the envelope of a natural background, and 3) a natural chorus of three other birdsongs. Within each noise category, the level of the foreground birdsong was modulated relative to the noise, giving a variety of target-to-masker ratios (TMRs). Unsurprisingly, across noise conditions response patterns became progressively dissimilar to the unmasked birdsong with decreasing TMR, indicative of the obfuscation of foreground features by increasing noise sound levels. More surprisingly, responses to the target were interfered with in two ways when masked: the addition of spikes in the gaps between foreground song syllables and the reduction of spikes during song syllables (Figure 1.15). The natural bird chorus background was found to reduce spiking during target syllables in all TMR conditions.

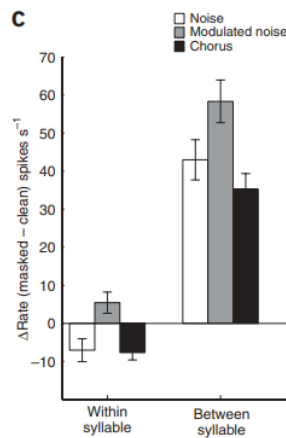


Figure 1.15: **Mean firing rate differences during and between target syllables.** Differences in firing rate (masked – unmasked response) for the three noise conditions both during and in the gaps between the target song syllables. Negative values show the reduction of spikes while positive values show the addition of spikes (Narayan et al., 2007).

Further, subjects were behaviorally tested on their ability to discriminate the target birdsong in the competing noise. Target identification accuracy degraded in parallel with decreasing TMR. Neural discrimination measured by fine timing was more consistent with behavioral performance compared to overall firing rate, suggesting that foreground discrimination fails when neurons are unable to precisely encode the timing of the target due to competing noise. Similarly, another study performs this general paradigm by recording responses from the primary auditory cortex of anesthetized cats during the presentation of bird chirps in either a clean condition or with their natural echoes or background noise from the recording. This study identified a population of neurons where responses to the



background/foreground together resembled the background more closely than the clean chirp (Bar-Yosef and Nelken, 2007).

These results challenge notions of ubiquitous noise-invariance by providing evidence for the single-unit representation of natural background noise even in the presence of a salient target. Still, the evidence remains unclear to what extent natural background noise is represented under fully natural and ethological conditions and remains to be fully explored. The work presented in this dissertation aims to more thoroughly address this gap with the goal of (1) better categorizing natural background/foreground response interactions at early stages of the auditory cortex, (2) determining the salient features of natural sounds that affect these interactions, and (3) how these responses are affected by training on a streaming task.

### ***Ferrets***

An animal model well-suited to addressing these goals is essential. Therefore, in most of the results described in this dissertation I used the ferret (*Mustela putorius furo*) as an animal model. Ferrets, though evolutionarily distant from humans, are an ideal model for studying higher-order auditory representations in part because they have a robust behavioral repertoire of auditory-dependent tasks such as tone detection (Fritz et al., 2003), sound discrimination (David et al., 2012; Heller et al., 2023), pattern recognition (Saderi et al., 2020), sound localization (Bajo et al., 2010), and auditory streaming (Ma et al., 2010). These complex behaviors are thought to be facilitated by a well-defined auditory hierarchy within cortex (Bizley and Cohen, 2013). Ferrets, unlike more common rodent animal models, anatomically have a well-defined auditory hierarchy within cortex (Figure 1.16) (Bizley et al., 2005; Bandyopadhyay et al., 2010; Hackett, 2011). Taken together and paired with the accessibility of *in vivo* recording methodologies, ferrets are well-suited to study higher-order auditory representations of natural sounds.

### ***Marmosets***

Given the surprising nature of some of our more foundational results, we hypothesized that discrepancies between our results in ferrets and established theories of auditory streaming in humans could be a result of the relatively extensive evolutionary distance between humans and ferrets. As such, we chose to replicate select results in the common marmoset (*Callithrix jacchus*), a new world monkey with extensive vocal communication behaviors and an AC similar in structure to humans (Figures 1.16) (de la Mothe et al., 2006; Slee and David, 2015; Eliades and Tsunada, 2019). Evolutionarily more proximal to humans, marmosets have an auditory system with anatomical and functional similarities to humans and share a similar hearing range and auditory perceptual abilities, including the ability to process

the harmonic structure of natural sounds, a key grouping feature used by humans (Bendor and Wang, 2005; Osmanski and Wang, 2011; Song et al., 2016; Feng and Wang, 2017).

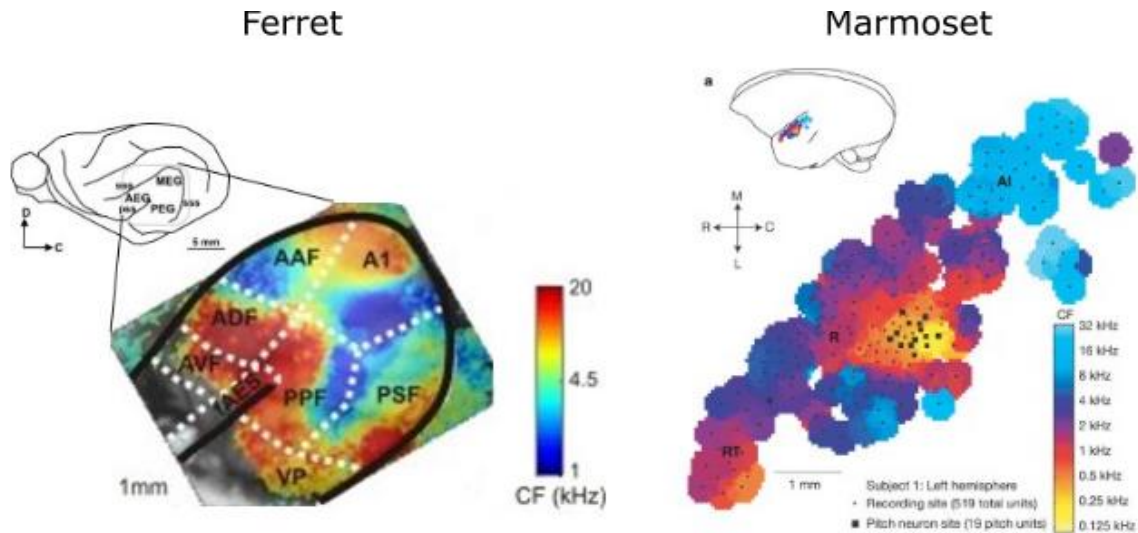


Figure 1.16: **Anatomical diagrams of AC organization and tonotopy in ferrets and marmosets.** *Left*, Diagram of ferret AC. Modified from (Bizley and King, 2009). *Right*, Diagram of marmoset AC. Reproduced from (Bendor and Wang, 2005).

## 1.2.5 Single-unit representations and auditory coding

In the previous sections reviewing the auditory streaming literature, I have made a point to emphasize the scale on which different methodologies allow us to view the problem of auditory streaming. For example, psychoacoustic studies allow us to choose and tailor sound stimuli to observe the perceptual effect these manipulations have on streaming, revealing functional details of the phenomenon. These studies can be complemented by population level electrophysiology recordings like EEG or fMRI which pull back the curtain on the brain to reveal, on a broad scale, computations performed during streaming. In animal models, we zoomed in further using invasive recording techniques which allow neural activity at the ensemble or individual level to be read out to determine more local computations.

The work in this dissertation will look at what information about overlapping natural sound stimuli are represented in neurons at the single-unit level across AC. That is, when sounds fitting into our broad background/foreground binary are played at the same time, which features of each sound category are encoded in the responses of individual neurons. As such, an important final step before discussing those results is a more detailed discussion of auditory coding and how it is modeled in AC neurons.

### ***Tuning properties***

In describing the anatomy of the auditory hierarchy in Section 1.2.1, we touched upon how neurons in different regions of auditory cortex possess unique tuning properties. More specifically, these areas have been well described in ferrets where the primary auditory cortex (A1) has sharper frequency tuning than secondary regions like the periectional gyrus (PEG), which tend to have shorter response latencies (Bizley et al., 2005). Still, even within a region different neurons can be tuned to very different spectrotemporal features of sounds.

This is because when we treat each neuron as an individual, we see these broad descriptions can be an oversimplification because each neuron actually responds to very precise spectral, temporal, and level characteristics. For example, two A1 neurons may be spectrally tuned so that they respond to a best frequency of 900 Hz, yet their temporal properties could vary in a way that one neuron fires throughout the duration of a 900 Hz tone while the other responds only to the onset and then is suppressed for the remainder. One of those same neurons may respond weakly to an 800 Hz tone not too far from its best frequency, yet the other may be completely suppressed by this spectral neighbor. How then can we most precisely and efficiently model these vast and complex tuning properties of auditory cortex neurons?

### ***The STRF***

The combination of spectral, temporal, and level properties that elicit spikes from a neuron define its receptive field, or the features to which a neuron is “tuned.” These unique sets of properties are summarized and modeled using the spectrotemporal receptive field (STRF, Fig. 1.17) (Aertsen and Johannesma, 1981). An STRF is a collection of weights estimated using a reverse correlation between neural responses and the spectrograms of the sounds that generate those responses. They describe neural tuning spectrally (what frequencies does the neuron respond to) and temporally (with what response latency must this spectral feature appear, considering the history of the sound) to predict whether a neuron will fire at any given moment when presented with a sound stimulus by weighing the presence of absence of the particular sound features.

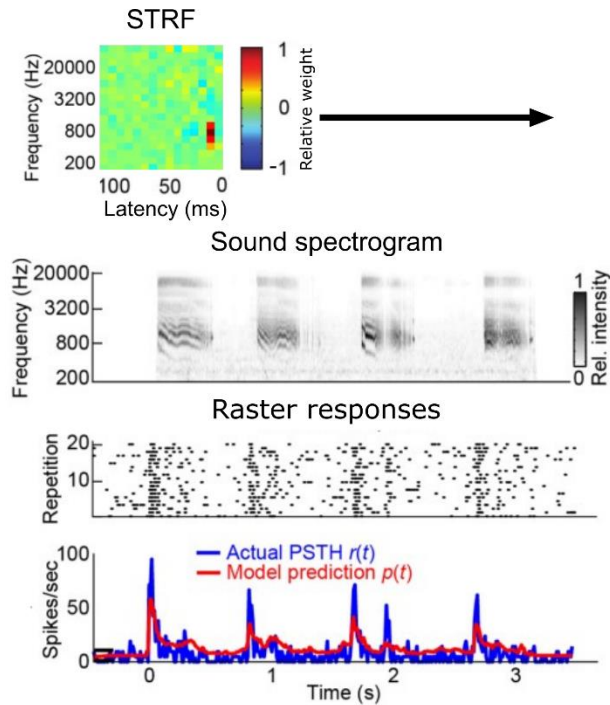


Figure 1.17: **Example spectrotemporal receptive field (STRF)**. At the top of the figure, an example STRF is shown for a single neuron. The spectral and temporal space is tiled by weights (represented by color) that describe whether a sound having power at that particular frequency and time lag will evoke a response by detecting the presence of tuned features. Weights closer to 1 are shown in red and indicate that a feature in that space will elicit a spike, while weights closer to -1 result in suppression by features in that space. The STRF shown, therefore, describes a neuron that is tuned to sounds around 800 Hz with a  $\sim 10$  ms response latency. Below the spectrogram is the actual raster response of a neuron, where each row is a repetition of that sound, and ticks indicate neural spiking. The neuron fires the most spikes at the onset of each bout of sound from the spectrogram above. Averaging across repetitions leads to the peristimulus time histogram (PSTH) at the bottom in blue, summarizing the raster and actual firing of the neuron. The red PSTH is the response predicted by summing the weights of the STRF at each time point as it slid over the spectrogram. At certain times, it will encounter moments where the sound onsets with power at 800 Hz which will result in the higher weighted areas of the model detecting this feature and predicting a neural response. Modified from (Thorson et al., 2015).

The STRF is a powerful model for describing more complex auditory tuning properties, but it is not without limitations. These arise from the linearity of the model—it describes a neuron as a static filter and feature detector. This assumption can be problematic for two reasons: (1) tuning properties of neurons become more complex and higher dimensional as we ascend the auditory hierarchy and (2) neural tuning is plastic and can be modified over long and short timescales. We will explore both and then relate them to how the STRF model may be applied to natural sounds and neural representations of sound mixtures during auditory streaming.

### ***Increased abstractions along the auditory hierarchy***

Along the auditory hierarchy from midbrain to secondary AC, spectral and temporal properties of the features describing neural tuning evolve as higher-level representations of sound stimuli are formed. For example, before reaching A1 temporal dynamics of neurons in subcortical regions of the auditory hierarchy respond with very rapid latency, as fast as 10 ms, yet in A1 latencies can be greater than 100 ms and still longer in secondary AC (Escabí and Read, 2003). Meanwhile, spectral tuning up through A1 shows similarly narrowband STRFs (Miller et al., 2002) in contrast to neurons in secondary AC which are more selective for broadband (Bizley et al., 2005) and complex spectral patterns (Rauschecker et al., 1995; Kikuchi et al., 2014). The change in neural tuning from rapid narrowband tone detector to a more leisurely detector of higher-level spectral patterns requires a greater number of spectral and temporal filters to capture these complex responses, thus becoming more non-linear and challenging to model using a single STRF (Atencio et al., 2009).

### ***Tuning in the auditory cortex is dynamic***

An STRF is static, but our perception is anything but static. The car alarm blaring outside your window may reflexively draw your attention and rapidly induce anxiety as you move to the window to determine if your car is in peril. You relax when you learn it wasn't your car and it gets disarmed. Shortly after the same car alarm goes off again, and although you may stir initially, the immediate urgency is gone. The third alarm within the hour, though the auditory stimulus is the same you may now react irked rather than concerned. In this case, over the course of less than an hour your perception and resulting reaction to a previously behaviorally salient sound has been modified by your updated experiential context. Similarly, at its core the Aesop's Fable, *The Boy Who Cried Wolf*, could be thought of as an early tale documenting plasticity in the auditory cortexes of a cohort of villagers as the behavioral salience of the warning call of "Wolf!" was extinguished as a result of the repeated decoupling of the auditory stimulus from the emotionally salient response.

This is to say that by existing in a dynamic world we constantly learn and relearn relationships between sounds and the potential for reward or punishment they bring. It has been well documented that learning the association between a pure tone target and a reward alters neural activity of AC with neurons both increasing their firing rate and reliability to tones around the target frequency (Diamond and Weinberger, 1986; Kisley and Gerstein, 2001; Hui et al., 2009). Similarly, neural tuning also reflects changes to reduce distractors (Schwartz and David, 2018).

Plasticity has been explicitly demonstrated to affect STRFs on rapid timescales corresponding to behavioral state (Fritz et al., 2003). Here, head-fixed ferrets were trained to withhold licking during the

presentation of broadband reference stimuli until a pure tone target was presented, at which point licking would result in a reward. Single units in AC were recorded and target tone frequency was selected based on initial passive STRF tuning of the cell. The same neuron was recorded during the behavioral detection task and again passively with no reward. STRFs were measured across these three trial blocks (passive-behaving-passive), revealing receptive field enhancement in auditory neurons at the target tone frequency during behavior and rapid reversion to the original tuning following behavior (Fig. 1.18). Neural encoding dynamically modified with the changing behavioral contexts and needs of the animal.

These referenced studies provide evidence for plasticity of neural activity to adapt to and enhance behaviorally salient pure tone stimuli while reducing features of noisy distractors. It remains unknown though what these receptive field adaptations may look like in the presence of comparable streaming-like tasks that require the discrimination of complex natural sounds, which evoke more complex patterns of activity in the AC than pure tones. This will be the major focus of the work presented in Chapter 3.

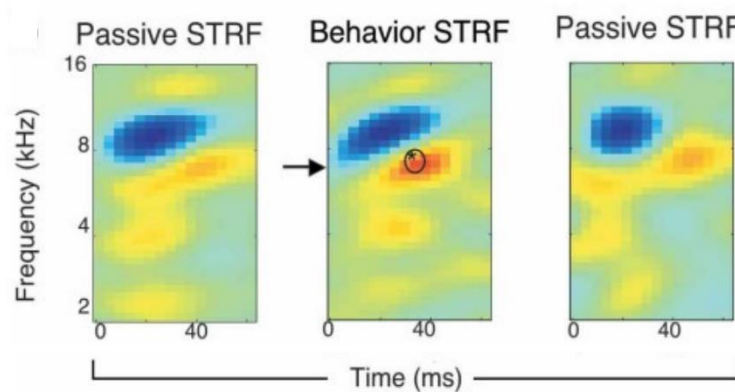


Figure 1.18: **Rapid, state-dependent receptive field plasticity in auditory cortex.** STRFs of a ferret auditory cortex neuron before, during, and after a pure tone detection task. Prior to behavior, the receptive neuron featured a large inhibitory region above 8 kHz and a small excitatory region around 7 kHz with a latency of 40 ms. With a target tone of 7 kHz (horizontal arrow) during behavior, the excitatory region at this frequency expands only to revert to approximately its original size in a subsequent passive recording taken after the behavioral task had ended. Reproduced from (Fritz et al., 2003).

### ***How do STRFs hold up in the presence of complex natural sounds or background noise?***

The STRF of a neuron is estimated as the reverse correlation of its activity and a snapshot of the stimulus spectrogram at that moment. To estimate an accurate STRF, neural responses from a large amount and diversity of stimuli probing different features is ideal. A classic method of estimating STRFs, therefore, is by using efficient synthetic sound stimuli like temporally orthogonal ripple combinations (TORCs) (Klein et al., 2000) which have complex and dynamically varying properties that densely sample a wide range of spectrotemporal relationships to determine tuning.

As mentioned throughout this introduction, the work presented in this dissertation will focus entirely on neural representations of natural sounds. Previously we discussed that throughout the primary and secondary regions of the AC the receptive fields of neurons become more complex and abstract, exemplified by the emergence of neurons tuned to specific categorizations of sounds like conspecific vocalizations (Montes-Lourido et al., 2021). At the same time, as these high-level auditory neurons become increasingly complex in their receptive fields, they no longer are activated by the relatively simple, synthetic stimuli traditionally used to estimate STRFs, including TORCs (Theunissen et al., 2000). Fortunately, STRFs also can be estimated using natural sounds (Theunissen et al., 2001), but require extensive and diverse natural sound exemplars to effectively and reliably sample the necessary spectrotemporal space.

In the discussion of the future directions of my work that will take place in Chapter 4, I will overview preliminary results weighing performance benefits of using a linear model like an STRF or deep learning approaches like convolutional neural networks (CNNs) to predict neural responses to natural sounds and combinations of natural sounds. For now, we must begin to ask how well an STRF can describe responses to not just natural sounds, but combinations of natural sounds. An STRF is a feature detector and a characteristic of auditory streaming-like overlapping sound configurations is the obfuscating and masking of prominent foreground features by noise (Fig. 1.5 and Section 1.2.4). Moreover, the masking of features by natural backgrounds can impose constantly changing statistical relationships, which can be a challenge from the perspective of a neuron whose receptive field is being described by a static filter.

Recent work with CNNs seeks to adapt the descriptive power of an STRF to the rapid timescales over which sound features and therefore the neurons representing them can change. The nonlinear receptive field model is referred to as the dynamic STRF (DSTRF), or an STRF whose spectrotemporal tuning is a composite of linear STRF functions for and depending on each changing moment of a stimulus (Keshishian et al., 2020). Looking at how these DSTRFs change over time reveals variations and nonlinearities in gain, time, and shape. In the context of auditory streaming, we've already seen how population level representations of behaviorally relevant foreground noise fluctuate as natural background noise changes or remains constant (Fig. 1.10) (Khalighinejad et al., 2019), a result of adaptation to background statistics. In Chapter 4 I will explore the potential of DSTRFs to show how the receptive fields of neurons adapt over time to background noise in streaming-like configurations.

### **1.2.6 How does my work fit into this context?**

We've now spent a considerable amount of time steeping ourselves in the world of auditory streaming from its conceptual underpinnings to how research has already defined and tackled the problem. We started the journey by describing how the sounds that reach our ears arise and, when they do, how our body and brain are equipped to transform fluctuations in air pressure into meaning—music, language, danger. We ran into a problem though because our auditory systems are equally equipped to transform fluctuations in air pressure of unwanted noise into a competing signal.

How these signals are parsed by the brain has been the subject of countless research endeavors, defining the perceptual principles that underlie our ability to stream through a more birds-eye phenomenological lens and more recently beginning to talk about more nuanced neuronal mechanisms of how this perception arises. The work presented in the subsequent chapters will aim to extend this extensive field of knowledge by discussing the specific, complex, and unexpected interactions that arise in single-unit representations of auditory cortex neurons in trained and untrained animals when they are presented with spectrotemporally complex mixtures of natural background and foreground sounds.

In chapter 2 I will discuss the basics of these interactions, defining them and then exhaustively exploring an unexpected result that foreground responses preferentially reduced when paired with natural background textures. This preferential reduction is robust to numerous manipulations and in a manner dependent on the natural statistics of both foreground and background. This data will deviate from established results and principles in higher-order human AC (some of which were reviewed above) that describe widespread noise-invariance as the genesis of noise-robust perception. I will instead suggest that background representation at earlier stages of auditory processing may serve as a reference by which network-wide activity can subtract the noisy response to enhance foregrounds at later stages. In this way, while these results maintain a level of unexpectedness, they also can fit nicely within the larger, existing narrative of progressive invariance to noise ascending the auditory hierarchy.

Next, in Chapter 3, I will extend these results by contextualizing the same paradigm from Chapter 2 within a behavioral task that explicitly controls the behavioral salience of certain sounds. Our review of studies of humans underscored the importance of attention and behavioral salience for modifying larger-scale representations of noise mixtures. As a result, specifically imposing behavioral salience allows me to directly test the effect on single-unit representations of target and distractor in early AC within the framework of a natural foreground/background sound contrast. The results here will extend those from Chapter 2 and show that changes in auditory tuning may also occur because of behavioral experience even in natural sounds.



Finally, Chapter 4 will round out this discussion in a more speculative manner, using the details and results of both studies to discuss where this work fits into the literature on auditory streaming as well as laying the next steps that follow from my research.

## 2. Single-unit representations of natural background/foreground contrasts in passively listening ferrets

In discussing the background of the auditory streaming of natural sounds, I posed several questions and emphasized gaps within the existing literature: How are dynamic, more behaviorally relevant natural sounds represented at the single-unit level of the auditory cortex during natural background noise? Are single-unit representations of these natural background/foreground mixtures comparable to past evidence of noise invariance in higher auditory areas in human auditory cortex? What distinguishes natural backgrounds/foregrounds to the auditory cortex and how do these features define their interactions? In the forthcoming results presented in this chapter, we seek to directly address these questions by recording single-unit responses to natural background/foreground sound pairings in primary and secondary auditory cortex of passively listening ferrets, assessing the interactions and their mechanisms.

The forthcoming manuscript is in press at *Journal of Neuroscience* and appears in full below.

---

### Reduced neural responses to natural foreground versus background sounds in auditory cortex

Abbreviated title: Cortical reduction of natural foreground sounds

Gregory R. Hamersky<sup>1,2</sup>, Luke A. Shaheen<sup>2</sup>, Mateo López Espejo<sup>1,2</sup>, Jereme C. Wingert<sup>2,3</sup>, Stephen V. David<sup>2,\*</sup>

<sup>1</sup>Neuroscience Graduate Program, Oregon Health and Science University, Portland, OR 97239, USA

<sup>2</sup>Oregon Hearing Research Center, Oregon Health and Science University, Portland, OR 97239, USA

<sup>3</sup>Behavioral and Systems Neuroscience Graduate Program, Oregon Health and Science University, Portland, OR 97239, USA

The authors would like to thank Sam Norman-Haignere for guidance on generation of model-matched synthetic stimuli and members of the David Lab for feedback on data analysis. This work was supported by NIH BRAIN Initiative Award R01EB028155 and NIH Research Project Grant R01DC014950 (S.V.D.)

## **Abstract**

In everyday hearing, listeners face the challenge of understanding behaviorally relevant foreground stimuli (speech, vocalizations) in complex backgrounds (environmental, mechanical noise). Prior studies have shown that high-order areas of human auditory cortex (AC) pre-attentively form an enhanced representation of foreground stimuli in the presence of background noise. This enhancement requires identifying and grouping the features that comprise the background so they can be removed from the foreground representation. To study the cortical computations supporting this process, we recorded single unit responses in AC of male and female ferrets during the presentation of concurrent natural sounds from these two categories. In contrast to expectations based on studies in high-order AC, single-unit responses to foreground sounds were strongly reduced relative to the paired background in primary and secondary AC. The degree of reduction could not be explained by a neuron's preference for the foreground or background stimulus in isolation but could be partially explained by spectro-temporal statistics that distinguish foreground and background categories. Responses to synthesized sounds with statistics either matched or randomized relative to natural sounds showed progressively decreased reduction of foreground responses as natural sound statistics were removed. These results challenge the expectation that cortical foreground representations emerge directly from a mixed representation in the auditory periphery. Instead, they suggest the early AC maintains a robust representation of background noise. Strong background representations may produce a distributed code, facilitating selection of foreground signals from a relatively small subpopulation of AC neurons at later processing stages.

## **Significance statement**

Perception of important sounds in a world cluttered with competing background noise requires the ability to segregate relevant and irrelevant sound sources. Most prior work investigating neural mechanisms of this background/foreground contrast has supported the theory that auditory cortex activity is largely invariant to background noise, consistent with evidence from behavioral studies. However, it remains unclear what information about background noise is represented at the single-unit level. Here, contrary to prevailing theories, we show a relative dominance of single-unit responses to natural background noise over responses to natural foreground sounds in ferret auditory cortex. A robust representation of background noise in early stages of auditory cortex may be necessary for grouping features into perceptual objects and selecting information from foreground signals for preferential representation in downstream brain areas.

## Introduction

When interacting with the world, listeners encounter auditory scenes containing dynamic, spectrally overlapping sounds. Accurate perception requires streaming, the grouping of sound features into representations of their distinct sources (Bregman, 1990; Griffiths and Warren, 2004). Because natural sounds are spectrally and temporally dynamic, they cannot be differentiated based on tonotopic channels inherited from the cochlea alone. Instead, signal grouping processes that support streaming must distinguish stimuli according to statistical regularities in the time and frequency domains (Bregman, 1990; Darwin, 1997; Nelken et al., 1999; Carlyon, 2004; McDermott, 2009; Winkler et al., 2009). Streaming is engaged by active, top-down processes, such as during selective attention to the voice of a single speaker (O’Sullivan et al., 2015), and pre-attentively, as during perception of behaviorally relevant stimuli in noise (Sussman et al., 2007; Mesgarani et al., 2014; Kell and McDermott, 2019). In both cases, the representation of one stimulus is enhanced relative to others. The computational processes by which neural responses to competing stimuli are identified and suppressed remain poorly understood.

Psychoacoustic experiments often use synthetic stimuli with parametric properties to precisely manipulate and probe boundaries of streaming (Bregman et al., 2000; Bizley and Cohen, 2013). These studies highlight the importance of spectral (Cusack and Roberts, 2000; Akeroyd et al., 2005; Popham et al., 2018; McPherson et al., 2022), temporal (Micheyl et al., 2010; Andreou et al., 2011; Shamma et al., 2011; Sollini et al., 2022), and spatial (Akeroyd et al., 2005; Middlebrooks and Onsan, 2012; Bizley and Cohen, 2013) sound statistics to successful streaming. Natural stimuli are more complex and dynamic, but similar principles can be applied to model their streaming (Mesgarani et al., 2014; Theunissen and Elie, 2014; Młynarski and McDermott, 2019).

One framework to study natural sound streaming is a background/foreground contrast. Behaviorally relevant foregrounds (e.g., speech and other vocalizations) are perceived preferentially over noisy backgrounds (wind, water, machinery, etc.) (Bregman, 1990). Consistent with subjective percepts, most sounds can be classified as foreground or background based on spectro-temporal properties (Singh and Theunissen, 2003; Kell and McDermott, 2019; Attias and Schreiner, 1997). Local field potential (LFP) and functional imaging (fMRI) data from human superior temporal gyrus (STG) show that activity evoked by foreground sounds is largely invariant to background noise, suggesting that higher order auditory cortex automatically streams foregrounds over backgrounds (Kell and McDermott, 2019; Khalighinejad et al., 2019). Less is known about representation of these competing stimuli by single units or at early stages of processing, such as primary auditory cortex (A1).

Studies using decoding analysis of single-unit neural data have also argued for noise-robust representations in A1 (Mesgarani et al., 2014; Moore et al., 2013; Rabinowitz et al., 2013). Foreground

stimuli can be reconstructed accurately from single-unit AC activity, even in the presence of static, synthetic noise. However, this approach does not exclude the possibility that neural responses also preserve information about backgrounds (Christison-Lagay and Cohen, 2014; Malone et al., 2017; Ni et al., 2017). Some studies have indicated that information about both backgrounds and foregrounds is represented at this earlier stage (Bar-Yosef and Nelken, 2007; Narayan et al., 2007; Kell and McDermott, 2017). Thus, it remains uncertain if foreground representations in A1 are enhanced similarly to human STG or if they remain mixed with representations of the background.

To measure background/foreground contrasts in early stages of AC, we recorded single-unit activity in A1 and secondary auditory cortex of awake, passively listening ferrets. Natural background and foreground sounds, with categorically distinct spectro-temporal statistics, were presented in isolation and concurrently. In contrast to results in STG, we report an unexpectedly strong reduction of A1 responses to the foreground relative to concurrent backgrounds. This reduction may result from a combination of subcortical feed-forward processing and central gain control mechanisms. These findings differ from the previous reports of ubiquitous noise-invariant representation of foreground stimuli across AC. Instead, additional processing of evoked activity in A1 is required for noise-robust representations in downstream areas.

## **Material and Methods**

### ***Surgical Procedures***

All procedures were approved by and performed in accordance with the Oregon Health & Science University Institutional Animal Care and Use Committee (IACUC) and conform to the standards of the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) and the United States Department of Agriculture (USDA). Seven young adult ferrets (six neutered, descended male, one spayed female) were obtained from a supplier (Marshall Farms). In each animal, sterile head-post implantation surgeries were performed under anesthesia to expose the skull over the auditory cortex (AC) and permit head-fixation during neurophysiology recordings. Surgeries were performed as previously described (Slee and David, 2015; Sadari et al., 2020; Heller et al., 2023). In brief, two stainless steel head posts were anchored along the midline using light-cured bone cement (Charisma, Kulzer). To improve implant stability, 8-10 stainless self-tapping set screws were mounted in the skull. Layers of bone cement were used to build the implant to a final shape amenable to neurophysiology and wound margin care, which included frequent cleaning and sterile bandaging. Following a two-week recovery period, animals were habituated to head-fixation and auditory stimulation.

Additional data was collected from AC of two spayed and neutered adult marmosets (one female, one male) obtained from the University of Utah. Surgical and experimental procedures performed on marmosets were the same as for ferrets.

### ***Acoustic Stimuli***

Digital acoustic signals were transformed to analog (National Instruments), amplified (Crown), and delivered through a free-field speaker (Manger) placed 80 cm from the animal's head, 0° elevation, and 30° contralateral to the recording hemisphere. Stimulation was controlled using custom MATLAB software (<https://bitbucket.org/lbhb/baphy>) and all experiments took place inside a custom double-walled sound-isolating chamber (Professional Model, Gretsch-Ken).

Auditory stimuli consisted of a pool of 70 natural sound excerpts, each 1 s in length, curated and segmented to contain power immediately after onset. Sounds were divided into two ethological categories, backgrounds (BGs) and foregrounds (FGs), based on simple statistics that, respectively, produce the percept of a sound texture or dynamic transient (Lesica and Grothe, 2008; Kell and McDermott, 2019). Sounds were root mean square (RMS) normalized to impose a 0 dB signal-to-noise ratio (SNR) between BG and FG categories for most experiments. Sound level was calibrated so that individual sounds were presented at 65 dB SPL. For the five animals analyzed throughout the study, all 1 s natural sound excerpts were presented in isolation to each recording site, and the 3-5 sounds from each category that evoked the largest average multi-unit response across the recording site were selected for experiments. In two additional animals, considered separately from our main analyses, a fixed set of 4 BGs and 4 FGs were presented across recording sites. Selected sounds were combinatorially paired across categories to create 9-25 unique BG/FG combinations. The full set of isolated and concurrent sounds was presented in random order 10-20 times per recording.

We tested several variations of BG/FG combinations:

***Dynamic sound onsets.*** To study dynamics of adaptation to concurrent stimulation (Fig. 4), we presented natural pairs in which either BG or FG was truncated to its 0.5-1s half while the paired sound was played in full (BG+hFG or hBG+FG, Fig. 4A, 4E). FG sounds used in dynamic conditions were selected so that they contained energy both at 0 and 0.5 s onsets.

***Binaural stimulation.*** In our default experimental configuration, all sounds were presented from a single speaker 30° contralateral to the recorded brain hemisphere (*contra*BG/*contra*FG). For binaural stimulation, we added a second speaker 30° ipsilateral to the recording hemisphere. In these recordings,

BG and FG positions were varied, defining three additional spatial configurations: *ipsi*BG/*contra*FG, *contra*BG/*ipsi*FG, and *ipsi*BG/*ipsi*FG.

Variable SNR. In a subset of experiments, we varied FG RMS power relative to BG by 5 dB and 10 dB to produce instances where FG was relatively louder (i.e., greater SNR) than BG.

### ***Neurophysiology***

To prepare for neurophysiological recordings, a small craniotomy (0.5-1mm) was opened over AC. Recording sites were targeted based on tonotopic maps and superficial skull landmarks (Bizley et al., 2005; Atiani et al., 2014) identified during implantation surgery. Initially, tungsten microelectrodes (FHC, 1-5M $\Omega$ ) were inserted into the craniotomy to characterize tuning and response latency. Short latency responses and tonotopically organized frequency selectivity across multiple penetrations defined the location of primary auditory cortex (A1) (Bizley et al., 2005), whereas secondary auditory cortex (posterior ectosylvian gyrus, PEG) was identified as the field ventrolateral to A1. The border between A1 and PEG was identified from the low-frequency reversal of the tonotopic gradient.

Once a cortical map was established, subsequent recordings were performed using two different electrode configurations. Experiments in animals 1-3 used 64-channel silicon electrode arrays, which spanned 1.05mm of cortical depth (Du et al., 2011). Experiments in animals 4-7 recorded from 960-channel Neuropixels probes (Jun et al., 2017). Typically, about 150 of the 384 active channels spanned the depth of AC, as determined by current source density analysis (see below). Data were amplified (RHD 128-channel headstage, Intan Technologies; Neuropixels headstage, IMEC), digitized at 30 KHz (Open Ephys) (Siegle et al., 2017), and saved to disk for further analysis. Spikes were sorted offline using Kilosort2 (<https://github.com/MouseLand/Kilosort2>) (Pachitariu et al., 2016), with spike sorting results manually curated in phy (<https://github.com/cortex-lab/phy>). A contamination percentage was computed by measuring the cluster isolation for each sorted and curated spike cluster, which was classified as a single unit if contamination percentage was less than or equal to 5%. Clusters above 5% were classified as multi-unit and excluded from analysis.

Neurophysiology recordings were performed in animals in a passive state while head fixed and unanesthetized, with sessions typically lasting 4-6 h. During recording sessions, a video camera and LFP were used to monitor the animal's state. If an animal fell asleep or displayed signs of stress, the recording was paused to awaken the animal or provide resolution.

### ***Inclusion Criteria***



Evoked activity was measured using the peri-stimulus time histogram (PSTH) response, averaged across 10-20 sound repetitions and sampled at 100 Hz. While 100 Hz sampling captures most sound evoked activity, we repeated the core weight gain analyses on the data sampled at 50 and 10 Hz. Results were indistinguishable from those measured with a sampling rate of 100 Hz, and thus we used this rate for all subsequent analyses.

To ensure only sound-responsive units were included in analyses, we calculated a signal-to-noise ratio for each stimulus/neuron pair based on the ratio of the PSTH response to the standard deviation of the response across repetitions (Fritz et al., 2003). Stimulus-neuron pairs with  $SNR \geq 0.12$  were considered sound-responsive and included for analysis (Fig. 3A). Neuron/stimulus pairs where responses to both BG and FG in isolation exceeded SNR threshold were categorized as responsive to both sounds (BG<sup>+</sup>/FG<sup>+</sup> - A1: 9,157/28,728, 31.9%, PEG: 3,926/14,770, 26.6%) and comprise the dataset used in most analyses. Other stimulus/neuron pairs were categorized as responsive to only one sound (BG<sup>0</sup>/FG<sup>+</sup> - A1: 3,421/28,728, 11.9%, PEG: 1,608/14,770, 10.9%; BG<sup>+</sup>/FG<sup>0</sup> - A1: 2,189/28,728, 7.6%, PEG: 1,077/14,770, 7.3%) or unresponsive (BG<sup>0</sup>/FG<sup>0</sup> - A1: 13,961/28,728, 48.6%, PEG: 8,159/14,770, 55.2%). Unresponsive neuron/sound pairs were excluded from all analyses.

Within responsive neuron/sound pairs, outlier instances for which BG or FG weights were less than -0.5 and greater than 2 were also excluded (A1: 8/9,157, PEG: 5/3,926). Further, only instances where the linear model fit well ( $r \geq 0.4$ , A1: 7,144/9,149, 78.1%, PEG: 2,656/3,921, 67.7%) were included. Among the neuron/sound pairs excluded based on model fit accuracy, response weights trended with the  $r \geq 0.4$  data.

### ***Stimulus Statistics***

To calculate sound statistics, waveforms were loaded from .wav files and normalized to have a variance of 1. Spectrograms were generated using gammatone filters with 10 ms time bins and 48 frequency channels, log-spaced 0.1-24 kHz (Katsiamis et al., 2007).

Bandwidth quantified the range of frequencies that contained the majority of power in the spectrogram. A power spectrum was calculated by averaging each spectrogram over time and computing the cumulative sum across frequency. Bandwidth was determined by identifying the frequency range, in octaves, between 15% and 85% of the total.

Spectral correlation described how closely power across spectral bands co-varies. The correlation coefficient (Pearson's R) was computed over time between each pair of frequencies in the calculated bandwidth range and averaged across frequency pairs.

*Temporal variance* was quantified by calculating variance over time in each spectral channel in the calculated bandwidth range and averaging across frequencies. This statistic is also referred to as “temporal non-stationariness” (Khalighinejad et al., 2019). Greater deviations indicate higher temporal variance, or a more dynamic and transient sound.

*Spectral overlap* was the fraction of a sound's bandwidth that overlapped the bandwidth of a concurrently presented sound. Note that because individual sounds varied in bandwidth, this metric is not commutative, as the overlap of A with B is not the same as the overlap of B with A.

### ***Synthetic Sounds***

We generated model-matched, synthetic BG and FG sounds using a published MATLAB toolbox (Norman-Haignere and McDermott, 2018). Four synthetic conditions were synthesized so that their spectral/temporal modulation statistics were matched to the original sound or random: 1. Preserved spectral and temporal statistics, 2. Matched spectral and random temporal statistics, 3. Matched temporal and random spectral statistics, and 4. Random spectral and temporal statistics. Synthetic BGs and FGs were paired within each synthetic category, and presentations were interleaved with natural BG/FG pairs.

### ***Neural Tuning Analysis***

In a subset of experiments, we also recorded neural activity during the presentation of a large, diverse set of natural sounds, as previously described (593 unique, 1 s samples) (Pennington and David, 2023). We used this data to measure each neuron’s spectrotemporal receptive field (STRF) (Thorson et al., 2015) and then fit a two-dimensional Gabor function to the STRF (Qiu et al., 2003). The center of the Gabor fit on the spectral axis defined the neuron’s best frequency (BF). We used BF to select the corresponding channel from the spectrogram of the BG and FG stimuli presented to that neuron (spectrogram generated using gammatone filters with 10 ms time bins and 32 log-spaced frequency channels, spanning 0.2-20 kHz) (Katsiamis et al., 2007). The spectral SNR (in dB) for a BG/FG sound mixture at each neuron’s BF was calculated as the log-ratio of power in the BF frequency channel of the FG relative to that of the BG:

$$Spectral\ SNR = 20 * \log_{10} \left( \frac{\sum BF\ channel_{FG}}{\sum BF\ channel_{BG}} \right)$$

### ***Laminar Depth Analysis***

We used current source density analysis to classify units by cortical layer (1/3, supragranular; 4, granular; 5/6 infragranular). The local field potential (LFP) signal was generated by lowpass filtering either the raw signal from the 64-channel silicon probe or the LFP signal from the Neuropixel probe

below 250Hz using a zero-phase shift (filter-filter method) 4<sup>th</sup> order Butterworth filter, followed by down-sampling to 500Hz. A custom graphical interface was used to mark boundaries between layers based on features of average sound-evoked LFP traces sorted by electrode depth ([https://github.com/LBHB/laminar\\_tools](https://github.com/LBHB/laminar_tools)). Layer-specific features included the pattern of current source density (CSD) sinks and sources evoked by best frequency-centered broadband noise bursts. Patterns were selected to match auditory evoked CSD patterns seen in AC of multiple species (Maier et al., 2010; Schaefer et al., 2015; Davis et al., 2023; Mendoza-Halliday et al., 2023). Each unit was assigned a layer based on the boundaries above and below the channel where its spike had the largest amplitude.

### ***Spike Width Classification***

We classified neurons as narrow- and broad-spiking based on the average width of the waveform. Width was calculated as the time between the depolarization trough and the hyperpolarization peak (Trainito et al., 2019). The distribution of spike width across neurons was bimodal, and the categorization threshold was defined as the minimum between the bimodal peaks. Filtering properties differed between 64-channel probes and Neuropixels, thus the categorization threshold was defined as 0.35 ms and 0.375 ms, respectively.

### ***Statistical analysis***

For all pairwise statistical tests (Figs. 2, 3, and 5-8), we performed a Wilcoxon signed-rank test. Significance was determined at the  $\alpha = 0.05$  level. The number of neuron/sound pair combinations and animals for each comparison are listed in the main text or figure legends, as are exact p-values. For statistical tests for unpaired or across-area comparisons (Figs. 3, 5, and 8), we used a Mann-Whitney U rank test. Significance was determined at the  $\alpha = 0.05$  level. The number of neuron/sound pair combinations in each compared group were indicated in the main text or figure legends, along with exact p-values. Error bars for population average data were computed by jackknifing (Efron and Tibshirani, 1986). In Figs. 2 and 8, one-sample T-tests were performed to assess relative gain deviations from 0. Significance was determined at the  $\alpha = 0.05$  level.

To evaluate the relationship between PSTHs/sound statistics and our weighted metric of response reduction (Figs. 3 and 6), we performed a linear regression. Correlation coefficients and p-values are reported in the main text or figure legends.

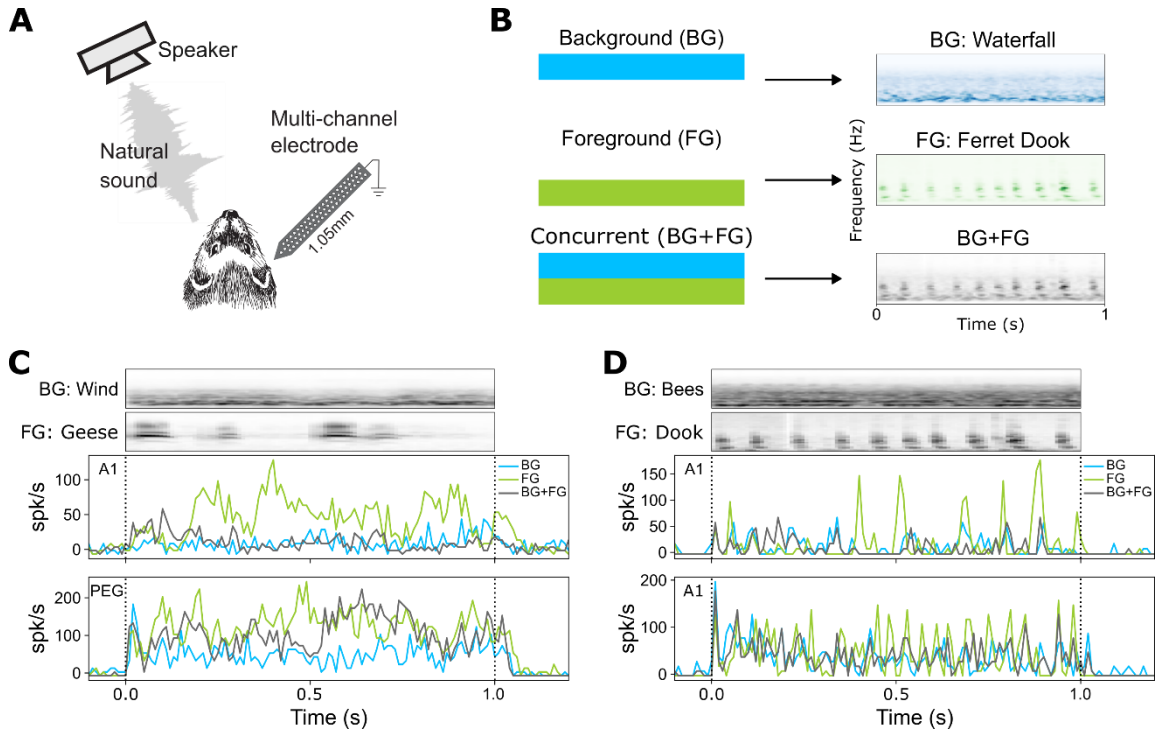
## Results

### Relative reduction of cortical responses to natural foreground sounds in the presence of concurrent backgrounds

To investigate how neurons in auditory cortex (AC) integrate information about concurrent natural sounds, we recorded single-unit activity from passively listening, head-fixed ferrets (Fig. 2.1A). Natural sound stimuli were drawn from two broad, ethologically relevant categories: background textures (BGs) and foreground transients (FGs). We used this BG/FG contrast based on previous work showing that this stimulus configuration produces streaming, with enhanced perception and cortical representation of the FGs over noisy BGs (Moore et al., 2013; Rabinowitz et al., 2013; Mesgarani et al., 2014; Kell and McDermott, 2019; Khalighinejad et al., 2019).

At the beginning of each experiment, a set of 29 BGs and 41 FGs (1 s duration) was presented in isolation, and the 3-5 sounds from each category that evoked the largest average multi-unit response were used in the subsequent recordings. Pairs of the selected BG and FG sounds were presented both individually and simultaneously (BG+FG, Fig. 2.1B). We recorded 1,191 auditory-responsive units in A1 (2,698 units total) and 601 auditory-responsive units in periectosylvian gyrus (PEG), a secondary field of auditory cortex (1,591 units total). Multiple BG/FG pairs were presented during each experiment. Data was selected based on whether a neuron responded to at least one BG or FG sound in the pair, leading to a total of 14,767/28,728 (51.4%) responsive neuron/sound pairs in A1 and 6,611/14,770 (44.8%) responsive pairs in PEG (see *Methods*). Peristimulus time histogram (PSTH) responses to each stimulus were computed by averaging across 10-20 repetitions to provide a measure of time-varying spike rate (Fig. 2.1C, D).

To gain a basic understanding of how AC neurons respond to concurrent natural stimulus pairs, we first evaluated response linearity. We compared the evoked PSTH response to each concurrent BG+FG stimulus to the sum of responses to the BG and FG stimuli in isolation. In both A1 and PEG, evoked responses to BG+FG combinations were consistently lower than the sum of the responses to same sounds in isolation. This global reduction of BG+FG responses was observed across most recorded units, consistent with previous studies of paired stimulus presentation (Kline et al., 2023).



**Figure 2.1.** Characterization of neural responses to concurrent natural background (BG) and foreground (FG) sounds in ferret auditory cortex (AC). *A*, Head-fixed ferrets were presented natural sound stimuli from a free-field speaker 30° contralateral to the recording hemisphere. Multi-channel microelectrode arrays recorded single-unit activity from primary (A1) or secondary (PEG) fields of AC. *B*, During recording, 1 s natural sound excerpts from two distinct, ethological categories—backgrounds (BGs) and foregrounds (FGs)—were presented in isolation (blue and green spectrograms, respectively) and concurrently (black spectrogram). *C*, Example PSTH responses to the same BG/FG sound pairing from units in A1 (*upper*) and PEG (*lower*). BG and FG responses are shown in blue and green, respectively, and the actual BG+FG response is shown in black. *D*, Example PSTH responses to the same BG/FG sound pairing from different units from the same recording site in A1. Color scheme as in *C*.

Given the observation of overall nonlinear response reduction, our next question was how the component BG and FG stimuli contribute to the concurrent BG+FG response. Patterns of reduction were heterogeneous but tended to follow two broad patterns: invariance to one of the stimuli (Fig. 2.1C, D, *upper*) or a combination of responses to the individual BG and FG stimuli (Fig. 2.1C, D, *lower*). To quantify how each component contributed to the BG+FG response, we fit a model for the concurrent response as a linear weighted sum of the constituent responses:

$$R_{BG+FG}(t) = w_{BG}R_{BG}(t) + w_{FG}R_{FG}(t)$$

Weights were fit for each neuron and stimulus pair, minimizing the mean-squared error prediction of the actual BG+FG response (Fig. 2.2A). The gain model measures changes in response variance such that weights <1 indicate reduction relative to the component response. We compared weights between BG and

FG categories for each neuron and stimulus pair tested. For A1 neuron/stimulus pairs with reliable responses to both individual BG and FG sounds ( $BG^+/FG^+$ ,  $n = 9,157/28,728$ , Fig. 2.3A, see *Methods*), weights were categorically divergent and dominated by the BG. On average,  $w_{FG}$  (median =  $0.256 \pm 0.003$ ) was significantly lower than  $w_{BG}$  (median =  $0.571 \pm 0.004$ , Wilcoxon signed rank test,  $p < 10^{-9}$ ), indicating a strong preferential reduction of FG responses (Fig. 2.2B, C). Only neuron/stimulus pairs with a good model fit ( $r \geq 0.4$ , 78.1%) were considered in analyses, but the same trend was observed for units with worse model performance. In PEG, neuron/stimulus pairs in the  $BG^+/FG^+$  group ( $n = 3,926/14,770$ ) with  $r \geq 0.4$  (67.7%) showed a similarly strong preferential reduction of FG responses ( $w_{FG}$ : median =  $0.285 \pm 0.004$ ,  $w_{BG}$ : median =  $0.530 \pm 0.006$ , Wilcoxon signed rank test,  $p < 10^{-9}$ , Fig. 2.2D). Reduction of FG responses in A1 and PEG was seen across animals ( $n = 5$ ), and a pattern showing a relatively larger reduction of FG responses in A1 was observed in 4/5 animals included in our main analyses.

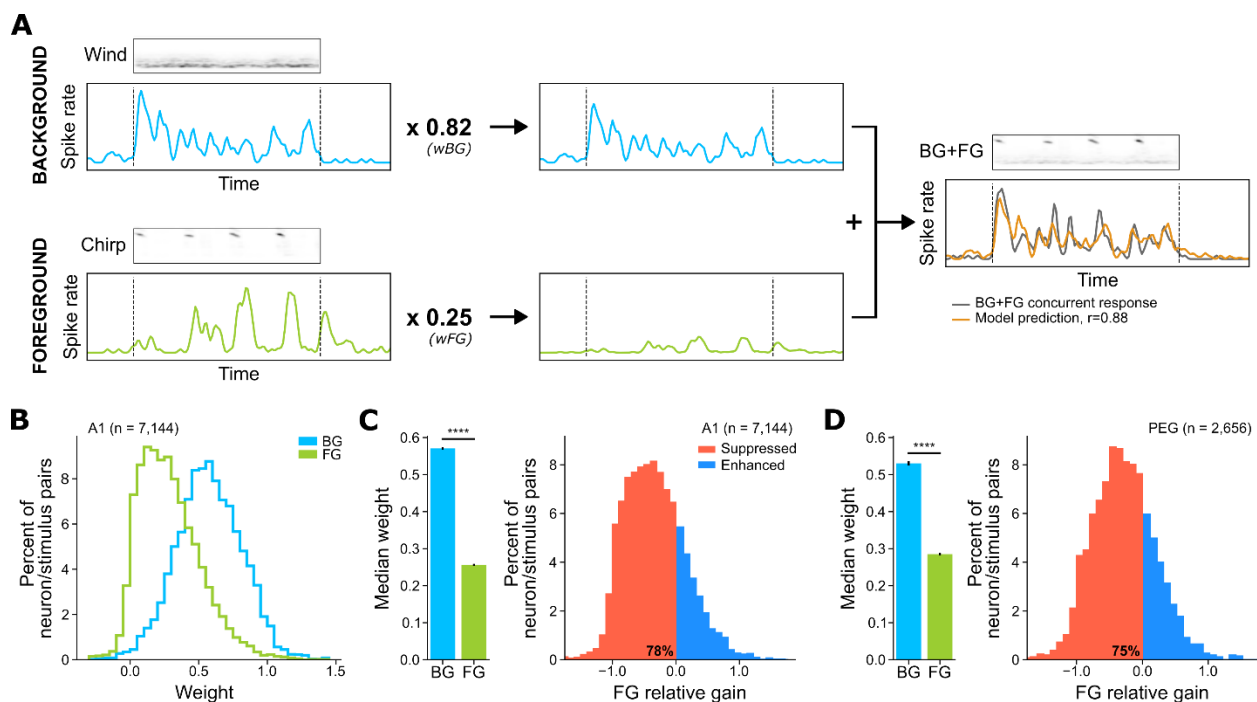
To verify that the weighted model recapitulated the nonlinear reduction in overall response reported above, we calculated the average weight,  $(w_{FG} + w_{BG}) / 2$ , for each neuron/sound pair. The distribution of average weights significantly correlated with level of reduction observed in the PSTH response to the concurrent sounds,  $R_{BG+FG} / (R_{BG} + R_{FG})$  (A1:  $r = 0.44$ ,  $p < 10^{-9}$ , Fig. 3B; PEG:  $r = 0.42$ ,  $p < 10^{-9}$ ). Thus, model weights provided a reasonable quantification of the overall and stimulus-specific responses.

Because stimuli were chosen to evoke activity at the current recording site, we considered the possibility that these effects may be specific only to sites tuned to the particular BG/FG set. In a separate cohort of animals ( $n = 2$ ), we paired a fixed set of BGs and FGs across all recording sites. In A1, neuron/stimulus pairs in the criterion  $BG^+/FG^+$  group (27.6%) with  $r \geq 0.4$  (73.8%) showed comparable weights ( $w_{FG}$ : median =  $0.187 \pm 0.008$ ,  $w_{BG}$ : median =  $0.591 \pm 0.006$ ,  $n = 1,564$ ) from those using sounds selected by evoked activity in our main analysis ( $w_{FG}$ : median =  $0.198 \pm 0.013$ , Mann-Whitney U rank test,  $p = 0.033$ ,  $w_{BG}$ : median =  $0.601 \pm 0.013$ , Mann-Whitney U rank test,  $p = 0.092$ ,  $n = 637$ ). Recordings from PEG showed fewer responsive neuron/stimulus pairs (11.2%), comparable fits where  $r \geq 0.4$  (67.0%), and a slightly reduced difference between FG and BG weights ( $w_{FG}$ : median =  $0.261 \pm 0.010$ ,  $w_{BG}$ : median =  $0.576 \pm 0.005$ ,  $n = 210$ ) relative to the main stimulus set ( $w_{FG}$ : median =  $0.175 \pm 0.008$ , Mann-Whitney U rank test,  $p = 0.005$ ,  $w_{BG}$ : median =  $0.628 \pm 0.017$ , Mann-Whitney U rank test,  $p = 1.55e-6$ ,  $n = 441$ ). These results confirm that the acoustic properties of the sound stimuli, independent of a recording site's tuning to the stimuli, determine the degree to which FG responses will be reduced.

Given that sounds from both categories were reduced relative to presentation in isolation (weight  $< 1$ ), we combined the weights into a single metric, FG relative gain ( $RG_{FG}$ ), to describe the relative contribution of FG versus BG to the BG+FG response:

$$RG_{FG} = \frac{w_{FG} - w_{BG}}{|w_{FG}| + |w_{BG}|}$$

Values of  $RG_{FG} < 0$  indicate a greater reduction of FG responses relative to BG, which we refer to as FG-specific response reduction.  $RG_{FG}$  values in A1 comprise a distribution centered significantly below zero ( $p < 10^{-9}$ , Fig. 2.2C, right). This systematic reduction of FG responses was not expected, given prior work indicating enhancement of responses to FG sounds in AC (Moore et al., 2013; Rabinowitz et al., 2013; Mesgarani et al., 2014; Kell and McDermott, 2019; Khalighinejad et al., 2019). Below, we describe several additional analyses to validate this result.



**Figure 2.2.** FG responses are preferentially reduced relative to BG responses. **A**, Diagram of the linear weighted model used to quantify the contribution of component BG and FG responses to concurrent BG+FG sound presentations. The PSTH response to each BG and FG sound in isolation was weighted and summed to minimize the mean-squared error prediction of the actual BG+FG response. **B**, Histograms show distribution of BG (blue) and FG (green) weights for all neuron/sound pair combinations in A1. **C**, Bars at left compare median weights ( $\pm$  jackknifed S.E. across neuron/sound pairs,  $n = 7,144$ , Wilcoxon signed-rank test,  $****p < 10^{-9}$ ). Histogram of FG relative gain ( $RG_{FG}$ ) for all A1 combinations. Negative  $RG_{FG}$  values (red) indicate neuron/sound pairs that show FG-specific response reduction ( $RG_{FG} < 0$  for 78%). **D**, (left) Median BG and FG weights in PEG ( $n = 2,656$ ,  $****p < 10^{-9}$ ) and (right) histogram of RG in PEG ( $RG_{FG} < 0$  for 75%), plotted as in C.

## Degree of FG response reduction does not depend on selectivity for component stimuli

One possible explanation for the reduction of FG responses and the dominance of BG responses could be that the balance of BG and FG weights is determined by the component sounds' alignment with the receptive field of a neuron. If this the case, the BG or FG sound evoking a stronger response in isolation would be expected to dominate the BG+FG mixture response, resulting in that sound being weighted more highly. To determine whether the relative reduction of FG responses could be predicted by the component sound responses, we calculated a z-scored response to each isolated sound by dividing the standard deviation of the PSTH during the sound-evoked window by the standard deviation computed across all stimuli. We compared the isolated BG or FG response to the corresponding weight (Fig. 2.3B) to determine if there was a relationship. Neurons in A1 showed a weak negative correlation between BG responsiveness and  $w_{BG}$ , opposite what would be expected if neural SNR predicted response weight ( $n = 7,144$ , linear regression,  $r = -0.11$ ,  $p < 10^{-9}$ , Fig. 2.3B, *left*). There was no relationship between FG responsiveness and  $w_{FG}$  (linear regression,  $r = 0.01$ ,  $p = 0.57$ , Fig. 2.3B, *right*). Similarly, PEG showed a small negative correlation between BG responsiveness and  $w_{BG}$  ( $n = 2,656$ , linear regression,  $r = -0.06$ ,  $p = 1.21e-3$ ). It also showed a small positive correlation between FG responsiveness and  $w_{FG}$  (linear regression,  $r = 0.08$ ,  $p = 4.28e-5$ ). Together, these results indicate that a neuron's tuning to a particular stimulus had a relatively small influence on its weighting in the response to stimulus mixtures. The relative influence of this factor is considered in the regression analysis, below.

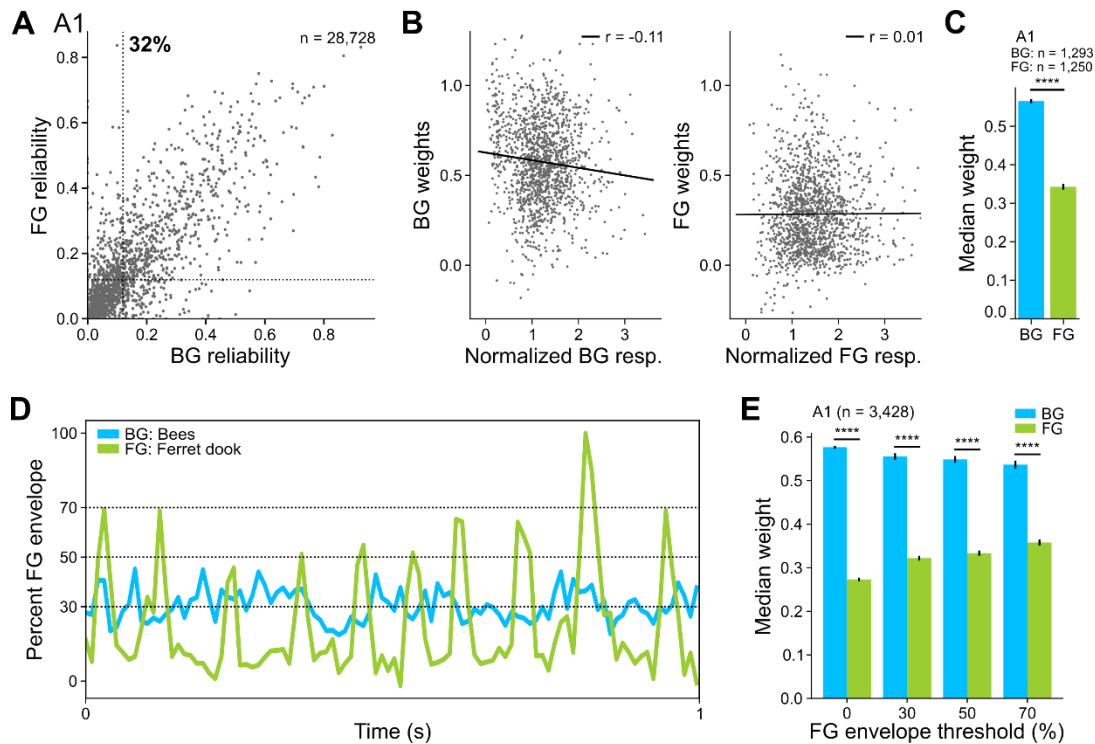
To consider a more extreme case where stimulus selectivity might impact relative weight, we analyzed the subset of stimulus pairs where only one sound, BG or FG, evoked a response in isolation. We defined the  $BG^0/FG^+$  and  $BG^+/FG^0$  groups as neuron/stimulus pairs that responded only to the FG or BG stimulus, respectively ( $n = 3,421/28,728$ , 11.9%,  $n = 2,189/28,728$ , 7.6%, Fig. 2.3A, see *Methods*). Comparison of  $w_{FG}$  from the  $BG^0/FG^+$  subset (median =  $0.343 \pm 0.008$ ) and  $w_{BG}$  from the  $BG^+/FG^0$  subset (median =  $0.565 \pm 0.006$ ) showed similar, significant ( $p < 10^{-9}$ ) FG response reduction in both A1 and PEG ( $w_{FG}$ : median =  $0.331 \pm 0.015$ ,  $w_{BG}$ : median =  $0.541 \pm 0.008$ ,  $p < 10^{-9}$ , Fig. 2.3C). The reduction of FG responses even in the absence of any BG-evoked response suggests that the reduction can be driven by subthreshold activity and thus does not simply reflect preferential responses for BG stimuli.

One other factor that might explain the unexpected reduction of FG responses was the relatively static nature of BGs compared to dynamic FGs. Though total stimulus power was matched between the paired BG and FG, moment to moment fluctuations in FG amplitude will change the instantaneous SNR of the FG relative to BG. To determine if the reduction of FG responses was driven simply by periods of low FG power, we incrementally omitted time bins of low FG power from the weight analysis (Fig. 2.3D). To select bins for omission, we calculated the FG sound envelope by averaging the spectrogram



across frequency channels. We then established thresholds whereby the envelope must exceed 30, 50, and 70% of maximum for a time bin to be included in the weight analysis. As the envelope threshold increases, the FG SNR increases such that the amount of FG-specific response reduction should become less extreme if instantaneous SNR contributed to the effect. However, across all thresholds,  $w_{BG}$  was significantly greater than  $w_{FG}$  ( $p < 10^{-9}$ , Fig. 2.3E), consistent with the condition in which no threshold was applied.

The lack of strong relationship between relative weight and a neuron's single-stimulus response (Fig. 2.3B) or instantaneous FG SNR (Fig. 2.3E) indicates that the selective reduction of FG responses is categorical and dependent on activity in the wider AC network. This observation prompted further investigation into features of the BG versus FG stimuli that can explain this unexpected result.



**Figure 2.3.** Minimal dependence of FG response reduction on neural tuning or instantaneous FG signal-to-noise ratio (SNR). **A**, Scatter plot compares reliability (signal power/noise power) of responses to BG and FG stimuli for each neuron/sound pair combination in A1. The threshold for sound responsiveness was set at 0.12 (dotted lines), labeling 32% as responsive to both BG and FG. **B**, Scatter plots compare the relationship between z-scored response to each stimulus in isolation and its weight in the combined response for each sound category (BG: linear regression,  $r = -0.11$ ,  $p < 10^{-9}$ , FG: linear regression,  $r = 0.01$ ,  $p = 0.57$ ). **C**, Median BG and FG weights for A1 neuron/sound pairs where the neuron is only responsive to BG or FG (median  $\pm$  jackknifed S.E. across neuron/sound pairs, Mann-Whitney U rank test, \*\*\*\* $p < 10^{-9}$ ). Weights are shown for the BG or FG stimulus that evoked a significant response when presented concurrently with a stimulus that did not produce a response, i.e., with zero weight. **D**, Example of FG

sound envelope threshold analysis. Blue and green lines show sound envelope (spectrogram averaged across the spectral axis) of BG and FG sounds, respectively. Dotted horizontal lines at 30, 50, and 70% of maximum FG indicate envelope thresholds used. Only time bins where the FG envelope exceeded threshold were included in the respective fit. *E*, Median BG and FG weights for neuron/sound pairs across differing sound envelope thresholds ( $\pm$  jackknifed S.E. across neuron/sound pairs, Mann-Whitney U rank test, \*\*\*\* $p < 10^{-9}$ ).

### Neural responses adapt rapidly to concurrent stimulus presentations

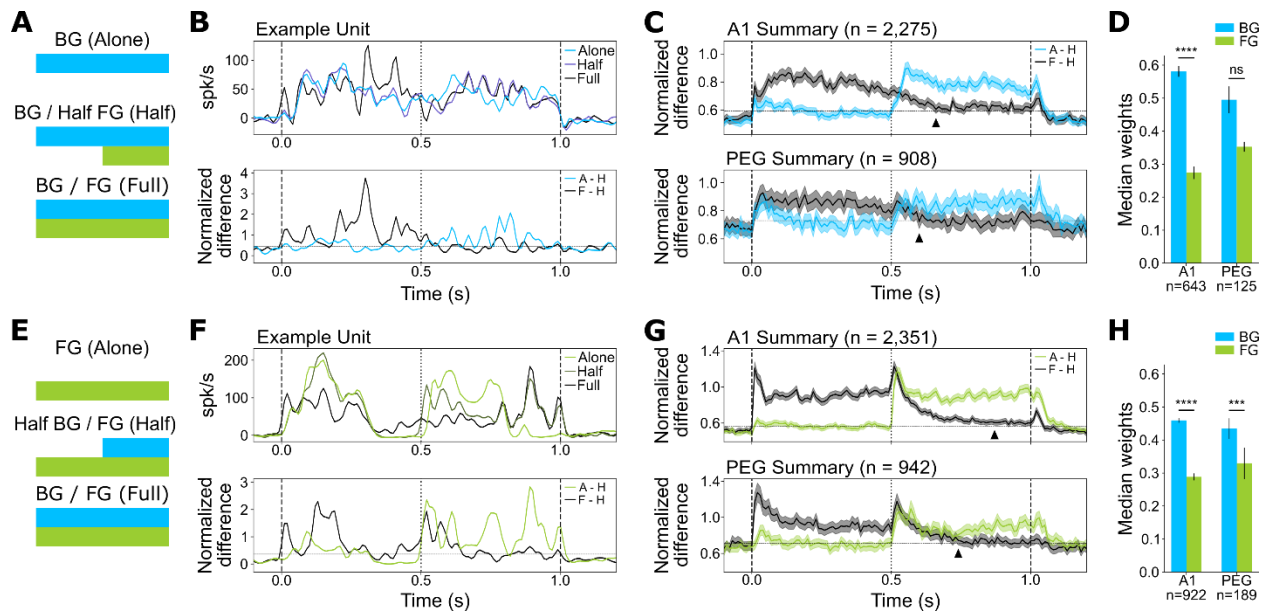
Theories of sound streaming suggest that the brain computes statistical regularities over time in the neural population response to group sound features into perceptual objects (Elhilali et al., 2009; McDermott and Simoncelli, 2011). To investigate the dynamics of BG+FG interactions, we measured the temporal window over which neural responses to a single sound adapt to the onset of a second sound. For a subset of recordings, we included stimulus instances where a BG played the entire 1 s stimulus duration and the paired FG began 0.5 s after BG onset (full BG + half FG; BG+hFG, Fig. 2.4A) or vice versa (late BG onset, hBG+FG, Fig. 2.4E). The late-onset stimuli were generated using the latter half of the full stimulus to allow direct comparison of responses to the second half of the standard BG+FG condition.

To measure dynamics of response to the onset of an interrupting FG, we compared the half-FG (BG+hFG) response to the full concurrent (BG+FG) response and to the BG alone response. Similarity of the BG+hFG response to the other conditions was computed as the difference in PSTH response, with the difference normalized by the standard deviation of the neurons' PSTH response across all stimuli. The first 0.5 s of the BG+hFG and BG alone conditions are identical; thus, their difference should be minimal and provide a baseline error, computed as the squared difference between PSTH responses. (Although stimuli during the 0-0.5 s window were identical for the BG+hFG and BG alone conditions, neural responses were variable trial-to-trial, resulting in a small, non-zero squared difference in the baseline. Similarly, response onsets tend to have higher spike rates. Combined with Poisson statistics of spiking, this onset produced a slight increase in baseline near time 0.) After the half-FG onset at 0.5 s, we expect the difference between the BG+hFG and BG+FG responses to decrease and converge to the baseline error. Example responses (Fig. 2.4B) demonstrate the timing of this transition in PSTH responses, which indicates how long the response takes to adapt to the appearance of the FG sound. The converse analysis was performed by comparing the response to an interrupting BG (hBG+FG) to both the BG+FG and FG alone responses (Fig. 2.4F).

Averaging the differential responses across all units and sound pairs for each condition allowed us to compute the average adaptation time following the onset of a second, concurrent stimulus. Because overall spike rate varied substantially between neurons, normalization prior to averaging provides a more accurate measure of change across the entire neural population rather than being dominated by high-firing

rate neurons. We defined the average adaptation time as the last time bin in which the response difference was significantly greater than the baseline error for three consecutive time bins ( $p < 0.05$ , jackknifed t-test). In the BG+hFG condition, adaptation in A1 took place more slowly (160 ms,  $n = 2,255$ ) than in PEG (100 ms,  $n = 908$ , Fig. 2.4C). Meanwhile, in the hBG+FG condition, which introduced the BG at 0.5 s, adaptation times were overall longer, but the same trend was observed between A1 (370 ms,  $n = 2,351$ ) and PEG (240 ms,  $n = 942$ , Fig. 2.4G). Thus, adaptation was slower following the introduction of BG sounds, but, as in the case of the introduction of FG sounds, it was faster in PEG than in A1.

Fitting the weight model to the 0.5-1 s window of the BG+hFG or hBG+FG conditions revealed adaptation to the initial, isolated stimulus influenced weights when the second stimulus appeared. In the BG+hFG condition, robust reduction of FG responses remained in A1 ( $n = 643$ ,  $p < 10^{-9}$ ), whereas PEG no longer showed a statistically significant difference between  $w_{BG}$  and  $w_{FG}$  ( $n = 125$ ,  $p = 0.374$ , Fig. 2.4D). Similarly, weights for the hBG+FG condition showed decreased FG response reduction in A1 and PEG, though both remained statistically distinct (A1:  $n = 922$ ,  $p < 10^{-9}$ , PEG:  $n = 189$ ,  $p = 6.83e-4$ , Fig. 2.4H).



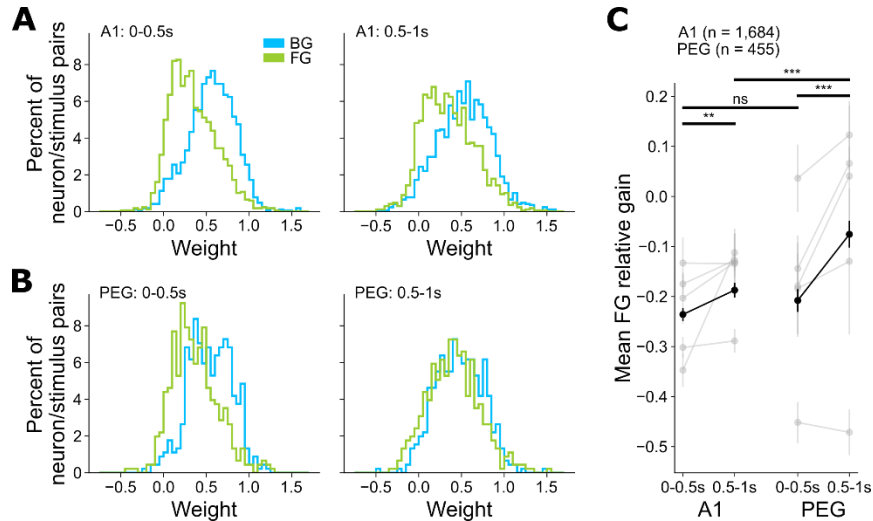
**Figure 2.4.** Response dynamics following concurrent sound onset differ between A1 and PEG. *A*, Schematic of BG+hFG condition, in which a truncated 0.5 s FG begins 0.5 s after BG onset. *B*, (*upper*) PSTH response of a single A1 neuron to BG+hFG, BG+FG, and BG alone stimuli. (*lower*) Difference in PSTH response between the BG+hFG condition and the BG+FG (black) and BG alone (blue). Values near zero indicate similar responses at the corresponding time point. *C*, Normalized response difference between BG+hFG and the two reference conditions, averaged across all neuron/sound pair combinations in A1 (*upper*) and PEG (*lower*). Filled triangles indicate the latest time point at which the response following introduction of FG at 0.5 s differs from the ongoing BG+FG response (A1: 160 ms, PEG: 100 ms).

Significant differences were measured relative to a noise floor, computed by averaging BG+hFG and BG alone responses over 0.5 s. **D**, Median BG and FG weights in the 0.5-1 s time window for BG+hFG stimuli (median  $\pm$  jackknifed S.E. across neuron/sound pairs, A1:  $n = 643$ , Wilcoxon signed-rank test, \*\*\*\* $p < 10^{-9}$ , PEG:  $n = 125$ , Wilcoxon signed-rank test,  $p = 0.374$ ). **E**, Schematic of hBG+FG condition, where the BG sound is introduced 0.5 s following FG onset. **F**, Example responses hBG+FG, BG+FG, and FG alone stimuli, plotted as in **B**. **G**, Average normalized response difference for hBG+FG stimuli, plotted as in **C** (A1: 370 ms, PEG: 240 ms). **H**, Median BG and FG weights in the 0.5-1 s time window for hBG+FG stimuli plotted as in **D** (median  $\pm$  jackknifed S.E. across neuron/sound pairs, A1:  $n = 922$ , Wilcoxon signed-rank test, \*\*\*\* $p < 10^{-9}$ , PEG:  $n = 189$ , Wilcoxon signed-rank test, \*\*\* $p = 6.83e-4$ ).

### Degree of foreground-specific response reduction decreases over time

Example neurons suggest FG responses in A1 are reduced throughout the duration of the BG+FG stimulus (Fig. 2.1C, *upper*) while in PEG the size of this effect can decrease over the course of the 1 s stimulus (Fig. 2.1C, *lower*). Based on our analysis of response dynamics (Fig. 2.4), we determined that BG+FG responses reach a steady state by 0.5 s. To compare relative FG response reduction before and after reaching steady state, we repeated the relative weight analysis (Fig. 2.2A) separately for the first and second 0.5 s halves of the BG+FG response.

A1 and PEG both showed significantly less reduction of FG responses in the second half of the stimulus (A1: Wilcoxon signed rank test,  $p = 7.36e-3$ ; PEG:  $p = 4.47e-7$ , Fig. 2.5).  $R_{FG}$  was not significantly different between A1 and PEG in the 0-0.5 s fit period ( $p = 0.096$ ). However, consistent with results showing more rapid adaptation dynamics in PEG (Fig. 2.4C, *G*), relative reduction of FG responses in PEG during the 0.5-1 s fit period was significantly less than in A1 ( $p = 9.91e-5$ ). Thus, both areas show selective reduction of FG responses following sound onset. After a period of adaptation, the amount of FG response reduction in PEG, which lies later in the auditory processing hierarchy, is smaller than in A1.



**Figure 2.5.** FG response reduction decreases after adaptation to a steady state. **A**, Histograms compare distributions of BG and FG weights in A1, fit using responses either during 0-0.5 s (*left*) or 0.5-1 s after onset (*right*). **B**, Distributions of BG and FG weights in PEG over the same time windows. **C**, Average FG relative gain ( $RG_{FG}$ ) between BG and FG weights, computed for each time window and cortical field. Data for the entire dataset is in black (mean  $\pm$  S.E.M. across area, Mann-Whitney U rank test,  $**p < 0.01$ , ns: not significant; across time windows, Wilcoxon signed-rank test,  $***p < 0.001$ ). Data from individual animals shown in gray ( $n = 5$ ).

### Distinct spectral and temporal sound statistics account for reduction of FG responses

Having established a categorical difference in response weights, we next sought to identify distinguishing statistical features of BG and FG stimuli that give rise to FG-specific response reduction. While the distinction between BGs and FGs can be intuitive—BG sounds typically contain less behaviorally relevant information (Rabinowitz et al., 2013)—they can also be distinguished by their statistical properties (Attias and Schreiner, 1996; Nelken et al., 1999; Singh and Theunissen, 2003; Lesica and Grothe, 2008; McDermott and Simoncelli, 2011). We evaluated the extent to which these distinguishing features can explain FG response reduction.

For each sound, we measured three spectro-temporal properties previously reported to distinguish BGs and FGs (detailed in *Methods*). Spectral correlation describes how closely power in different spectral bands co-varies (McPherson et al., 2022; Theunissen et al., 2000; Overath et al., 2008). Noisier sounds are typically less correlated due to the random nature of noise. As such, FG stimuli had greater spectral correlation than BG stimuli (BG:  $0.177 \pm 0.004$ , FG:  $0.488 \pm 0.014$ ,  $p = 1.65e-4$ , Fig. 2.6A). Temporal variance describes how transient or dynamic a sound is by computing variance in power across each frequency band over time and averaging across frequency (Khalighinejad et al., 2019). FG sounds tend to contain more transients (Lesica and Grothe, 2008) and thus have greater temporal variance than BGs (BG:  $0.073 \pm 0.004$ , FG:  $0.159 \pm 0.003$ ,  $p = 2.40e-7$ , Fig. 2.6C). Finally, bandwidth describes the frequency

range over which most of a sound's power resides. FG sounds tend to have narrower bandwidth than BG sounds (BG:  $3.32 \pm 0.18$ , FG:  $2.27 \pm 0.36$ ,  $p = 3.82e-3$ , Fig. 2.6E).

While these statistical properties differ between BGs and FGs on average, their values can vary widely across individual sounds within each category. We next compared how the magnitude of each property affects relative response gain during concurrent sound presentation. In the analysis above, we used FG relative gain ( $RG_{FG}$ ) to describe the relative contribution of FG versus BG to the BG+FG response. We define a complementary statistic, BG relative gain ( $RG_{BG}$ ), to describe the relative contribution of BG to the BG+FG response:

$$RG_{BG} = \frac{w_{BG} - w_{FG}}{|w_{FG}| + |w_{BG}|}$$

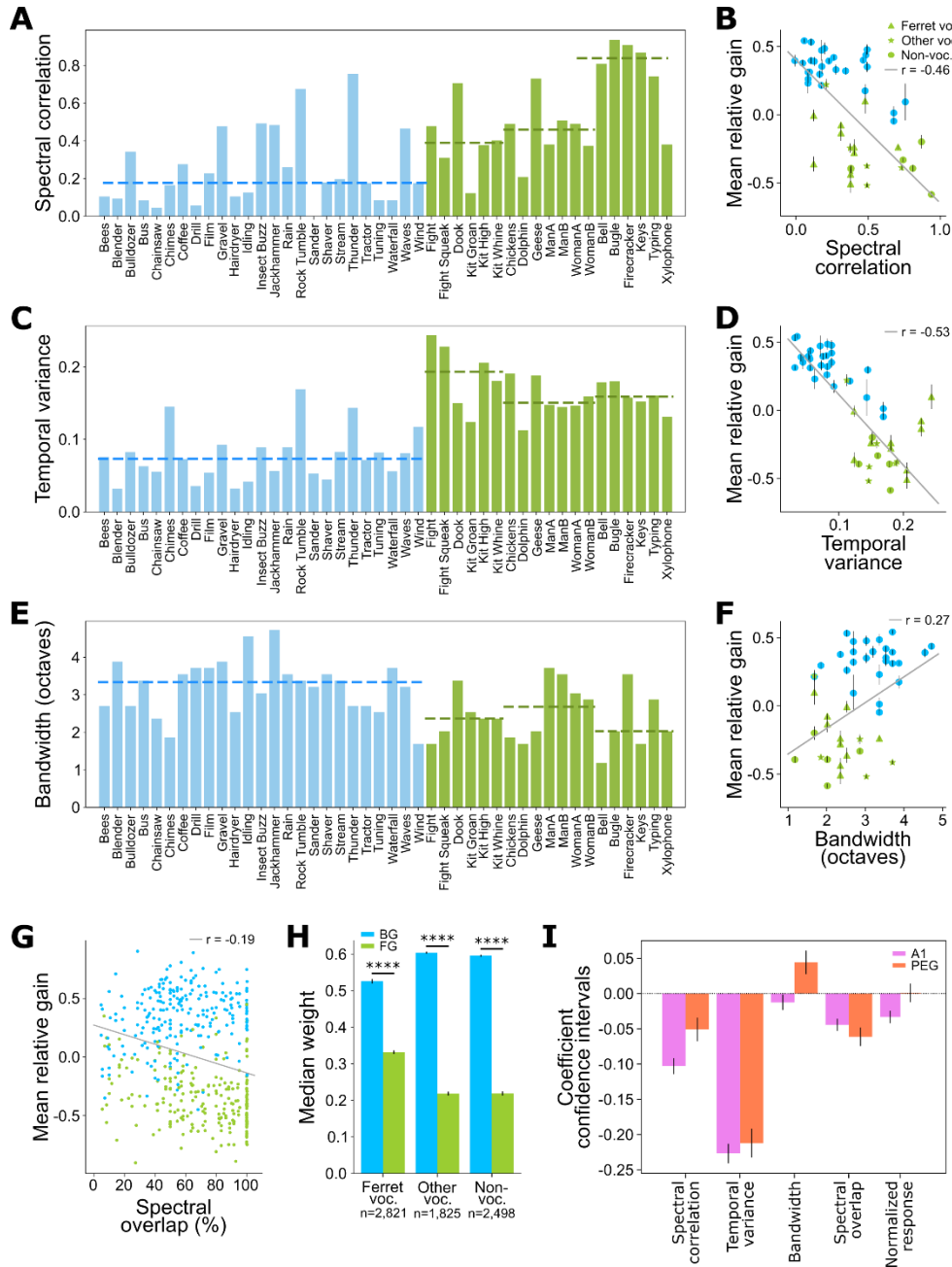
We can now describe the relative gain of any sound, BG or FG, using a single relative gain (RG) metric. As such,  $RG > 0$  indicates that the referenced BG or FG sound's response is relatively enhanced by a paired sound, and  $RG < 0$  indicates that the referenced sound's response is relatively reduced by a paired sound. We measured the relationship between each sound statistic and RG, averaging across neurons and paired stimuli. In A1, spectral correlation was negatively correlated with RG ( $r = -0.46$ ,  $p < 10^{-9}$ , Fig. 2.6B), temporal variance was negatively correlated with RG ( $r = -0.53$ ,  $p < 10^{-9}$ , Fig. 2.6D), and bandwidth was positively correlated with RG ( $r = 0.27$ ,  $p < 10^{-9}$ , Fig. 2.6F). Thus, differences in RG can be attributed to quantitative sound properties rather than the broad BG versus FG categorization. Similar relationships were observed in PEG (Spectral correlation:  $r = -0.36$ ,  $p < 10^{-9}$ ; Temporal variance:  $r = -0.52$ ,  $p < 10^{-9}$ ; Bandwidth:  $r = 0.33$ ,  $p < 10^{-9}$ ).

The properties described above characterize each sound in isolation. We also considered the possibility that the degree of overlap between sounds could explain their interaction. Spectral overlap, or the extent to which the spectral bandwidths of two sounds are matched, has been shown to affect the perception and encoding of concurrent sound sources (McDermott and Oxenham, 2008; Best et al., 2013). To calculate spectral overlap, we identified the range of frequencies used to measure bandwidth for each sound. Overlap of sound A with sound B is then defined as the percent of sound A's bandwidth that overlapped with sound B's bandwidth (note that this value is not commutative). In both A1 and PEG, there was a small negative correlation of spectral overlap with RG (A1:  $r = -0.19$ ,  $p < 10^{-9}$ , Fig. 2.6G; PEG:  $r = -0.20$ ,  $p < 10^{-9}$ ).

In addition to differences in spectro-temporal properties distinct between BG and FG sounds, we hypothesized that FG sub-categories contain inherent ethological salience to ferrets and thus may be represented differently. We divided FGs into three categories that might have different significance: 1.

ferret vocalizations, 2. vocalizations by other species, and 3. non-vocalizations (Fig. 2.6A, *green bars*). While all three of these categories were significantly reduced ( $p < 10^{-9}$ ), ferret vocalizations were the least reduced both in A1 ( $p < 10^{-9}$ , Fig. 2.6H) and PEG ( $p < 10^{-9}$ ), possibly reflecting their inherent behavioral relevance.

These observations together suggest that multiple sound statistics can explain the relative reduction of FG responses. At the same time, these properties can be correlated, as is the case of spectral bandwidth and spectral overlap, making it difficult to determine which ones actually account for the reduced response. To isolate their unique contributions, we performed a multivariate regression to predict RG based on a combination of sound statistics—spectral correlation, temporal variance, bandwidth, spectral overlap—and baseline individual response amplitude. Regression coefficients largely reflected analyses of individual statistics described above (Fig. 2.6I). Responses to sounds with either high temporal variance or high spectral correlation tended to be reduced. Spectral overlap and bandwidth had a significant yet relatively small effects. The relatively small effect of bandwidth was surprising, given prior results demonstrating that broadband, spectro-temporally dense sounds reduce AC neuron responses (Blake and Merzenich, 2002) and mediated by broadly-tuned lateral inhibition (Kato et al., 2017). Instead, the magnitude of response reduction appears to depend on tuning of network-level activity to higher-order stimulus features than bandwidth. Consistent with results showing weak or no relationship between single-stimulus response and weight (Fig. 2.3B, C), the regression indicates a weak effect of response strength in A1 and none in PEG. The broad effect of sound statistics on RG prompted further, direct investigation into the effect on RG of manipulating these natural sound statistics.



**Figure 2.6.** Distinct spectral and temporal sound statistics account for FG-specific response reduction. **A**, Spectral correlation measured for each BG (blue) and FG (green) sound. Horizontal dashed lines indicate mean across BGs and subsets of FGs. **B**, Scatter plot compares spectral correlation and relative gain for each natural sound in A1 (mean  $\pm$  S.E.M. per sound; linear regression,  $r = -0.46$ ,  $p < 10^{-9}$ ). Symbols indicate FG sub-category. **C**, Temporal variance of each BG and FG sound, plotted as in **A**. **D**, Scatter plot compares temporal variance versus relative gain ( $r = -0.53$ ,  $p < 10^{-9}$ ), plotted as in **B**. **E**, Bandwidth of each BG and FG sound, plotted as in **A**. **F**, Scatter plot of bandwidth versus relative gain ( $r = 0.27$ ,  $p < 10^{-9}$ ), plotted as in **B**. **G**, Scatter plot compares spectral overlap and relative gain for each BG/FG pairing (linear regression,  $r = -0.19$ ,  $p < 10^{-9}$ ). **H**, BG and FG weights, after FGs are grouped into vocalization sub-categories (median  $\pm$  jackknifed S.E. across neuron/sound pairs, Wilcoxon signed-rank test, \*\*\*\* $p < 10^{-9}$ ). **I**, Relative weight of each sound statistic's contribution to FG response reduction, measured by multivariate



linear regression. All variables were continuous. Responsiveness for individual BG or FG sounds were included as an input to control the influence of selectivity for the component sounds. Each input was normalized to have a variance of 1, permitting direct comparison of weights. Regression was performed independently for A1 and PEG.

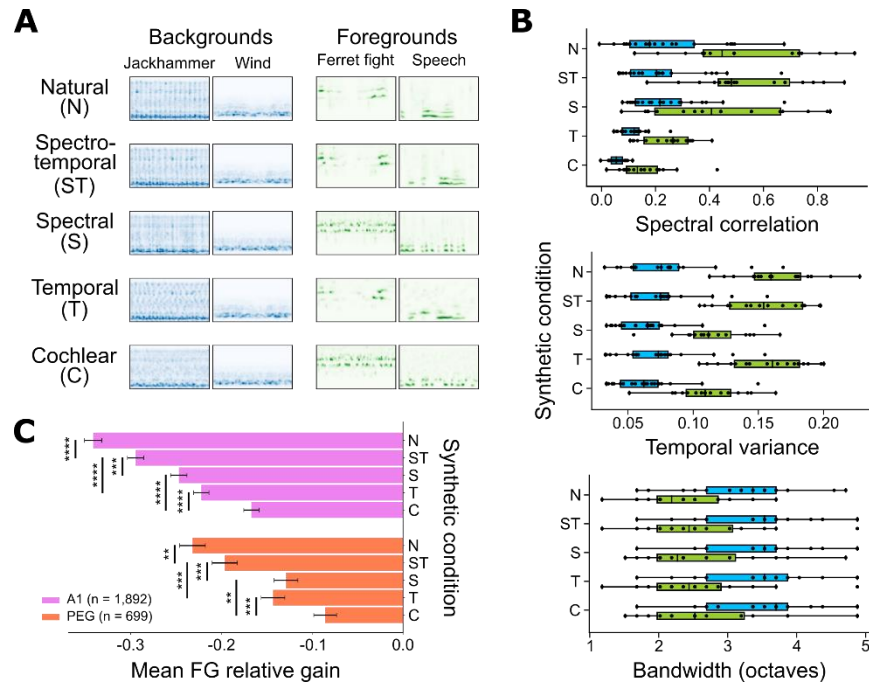
### **Synthetic stimuli confirm the role of natural spectro-temporal properties in preferential reduction of FG responses**

To directly test the role of spectro-temporal sound statistics on relative reduction of FG responses, we generated synthetic BG and FG sounds in which different combinations of temporal and/or spectral modulation features were selectively preserved (Norman-Haignere and McDermott, 2018). For a subset of recordings, synthetic stimuli were generated to match each natural BG/FG pair and presented on randomly interleaved trials. All stimuli were matched in their power spectrum (i.e., “cochlear”-level statistics). Four synthetic conditions were tested: 1. Preserved spectral and temporal statistics (spectro-temporal: ST), 2. Preserved spectral and random temporal statistics (spectral: S), 3. Preserved temporal and random spectral statistics (temporal: T), and 4. Random spectral and temporal statistics (cochlear: C, Fig. 2.7A). Pairs were always presented from the same natural/synthetic category.

We measured spectral correlation, temporal variance, and bandwidth across the different synthetic conditions (Fig. 2.7B). As expected, based on our analyses of the natural sounds (Fig. 2.6), spectral correlation between BGs and FGs were distinct in sounds generated with matched spectral statistics (ST, S). Meanwhile, FG stimuli with randomized spectral statistics (T, C) showed substantially decreased spectral correlation, showing a convergence toward BGs. Measurements of temporal variance followed a similar pattern, whereby FG and BG stimuli with natural temporal statistics (ST, T) were more distinct than those with randomized temporal statistics (S, C). Randomizing temporal statistics in FGs resulted in a much smaller decrease of temporal variance than the corresponding spectral statistics, likely a result of randomization disrupting but not completely abolishing transient temporal structures. Measurements of bandwidth remained unchanged across synthetic conditions, demonstrating that cochlear statistics were indeed conserved even when spectral and temporal statistics were randomized.

We next measured how synthetic sound pairs with randomized spectral and/or temporal statistics affected FG versus BG response weighting (Fig. 2.7C). There was a significant increase in FG relative gain between the original natural sounds and the spectro-temporal synthetic sounds, indicating that higher order statistics and relationships present in natural sounds contribute to the relative gain effect. The reduction in FG relative response became even smaller with the randomization of either natural spectral or temporal modulations. In the cochlear synthetic condition, where both spectral and temporal statistics were randomized, the difference in weighting was further reduced, although not eliminated. These persistent differences could result from residual temporal variance following temporal randomization or

from the presences of bandwidth differences that were conserved in the “cochlear”-level statistics (Fig 7B). The relatively strong effect of temporal variance in the regression analysis (Fig. 2.6J) suggests that the former may have the greater effect in this condition. Overall, both A1 and PEG showed similar successive decreases in FG response reduction, but with overall less reduction in PEG for each condition, consistent with comparisons between areas above (Fig. 2.2, 2.4, 2.5).



**Figure 2.7.** Sounds synthesized with randomized natural spectral and/or temporal properties successively decrease FG-specific response reduction in A1. **A**, Spectrograms of BG and FG natural sounds and four synthetic conditions, colored by category. Labels at left indicate the statistical properties (spectral, temporal modulations) preserved in each example. The cochlear condition lacks natural spectral and temporal modulation statistics. **B**, Whisker plots show distribution of BG and FG sound statistics—spectral correlation (*upper*), temporal variance (*middle*), and bandwidth (*lower*)—in synthetic conditions matching different spectral and/or temporal properties of the original natural sound. **C**, Average  $RG_{FG}$  for each synthetic condition in A1 and PEG (mean  $\pm$  S.E.M. across synthetic conditions, Wilcoxon signed-rank test,  $**p < 0.01$ ,  $***p < 0.001$ ,  $****p < 10^{-9}$ ).

### Spatial and signal-to-noise relationships between sounds influence FG-specific response reduction

Spatial location has been implicated as playing an ancillary role in streaming, operating largely to supplement monaural streaming cues like harmonicity and temporal onset (Shinn-Cunningham, 2005). Changing the spatial position of a sound source produces level differences at a single ear, which may also affect representation of sounds in mixtures. In the results described above, stimuli were all presented from a single location  $30^\circ$  contralateral to the recorded brain hemisphere (*contraBG/contraFG*). To explore the

effect of relative spatial location on FG responses to sound mixtures, we varied the location of each stimulus between the contralateral position and a second location 30° ipsilateral to the recorded brain hemisphere. This defined three additional spatial configurations: *ipsiBG/contraFG*, *contraBG/ipsiFG*, and *ipsiBG/ipsiFG*. Given the tendency of AC to preferentially encode contralateral stimuli (Middlebrooks and Pettigrew, 1981; King and Middlebrooks, 2010), we expected less reduction in FG response (i.e., less BG dominance) when the BG was presented ipsilaterally. Conversely, we expected greater reduction in FG response when the FG was presented ipsilaterally. Indeed, A1 and PEG both showed significant decrease in BG dominance (measured by  $R_{FG}$ ) in the *ipsiBG/contraFG* condition (A1:  $p < 10^{-9}$ , PEG:  $p < 10^{-9}$ ) and significant increases in BG dominance in the *contraBG/ipsiFG* condition (A1:  $p < 10^{-9}$ , PEG:  $p < 10^{-9}$ , Fig. 2.8A). The *ipsiBG/ipsiFG* condition showed a modest decrease in BG dominance in A1 ( $p = 0.019$ ) and no significant difference in PEG ( $p = 0.655$ ).

Varying sound location affected the relative loudness of sounds reaching each ear but also engaged differential circuits for spatial coding. To selectively test the effect of relative sound level on responses to the concurrent stimuli, we returned to the *contraBG/contraFG* configuration and instead varied the level of FG relative to BG successively from the original 0 dB SNR to 5 dB and 10 dB. Increasing SNR significantly and incrementally decreased FG-specific response reduction in A1 and PEG (Fig. 2.8B). Thus, while this preferential reduction only weakly depends on a neuron's responsiveness to individual BG and FG sounds (Fig. 2.6J), the degree does depend on overall input strength of each stimulus into the recorded brain area.

Effects of increasing FG SNR recapitulate the pattern of decreased FG response reduction when sound power envelope thresholds were imposed (Fig 3D). By fitting only to epochs above 30, 50, or 70% of the maximum FG strength, SNR was effectively higher, which resulted in incrementally smaller FG-specific reduction (Fig. 2.3E). These effects are also consistent with the *ipsiBG/contraFG* spatial condition, where the FG sound has higher effective SNR in the contralateral ear, which is preferentially represented in A1 (Stecker and Middlebrooks, 2003). The increased dominance of the BG in the *contraBG/ipsiFG* suggests that this trend would increase further for negative FG SNRs.

An important consideration in our experiments exploring relative sound level is that we always kept one stimulus at least at a level of 65 dB SPL, approximately 30 dB above ferret response threshold (Bizley et al., 2005). This sound level is advantageous as it evokes activity from a high proportion of neurons, but it remains possible that FG-specific response changes could differ at lower sound levels due to the non-monotonicity of some auditory cortical neurons. In non-monotonic neurons, it is possible that rather than having muted responses at lower sound levels, unique and novel interactions may arise. Non-monotonicity has been described as sparsely present in both primary and secondary areas of ferret

auditory cortex, with neurons in PEG tending towards non-monotonicity in greater abundance than A1 (Bizley et al., 2005). Thus, some of the effects reported here could differ if tested at lower sound levels.

### **Spectrotemporal interaction of natural sounds influence FG-specific response reduction**

While the average sound pressure level of the BG and FG sound waveforms was matched to have 0 dB SNR, mechanisms in the peripheral auditory system could influence the effective SNR of the combined spectrotemporal pattern driving the neural response. To consider this possibility, we measured relative energy of the two stimuli in the spectral band driving the neural response. For a subset of recordings, we collected a separate natural sound dataset that permitted estimation of a neuron's spectrotemporal receptive field and, from that, its best frequency (BF). We then computed a "spectral SNR" from the relative amplitude of FG and BG spectrograms in the BF channel (see *Methods*). Across this dataset, 61.0% of BG/FG mixtures had a negative spectral SNR (mean A1:  $-4.08 \pm 0.29$  dB,  $n = 2,835$ , PEG:  $-5.14 \pm 0.60$  dB,  $n = 664$ ). Thus, the BG often had greater relative power in the BG/FG spectrogram. In A1, spectral SNR was positively correlated with a  $FG_{RG}$  value, i.e., less FG response reduction (linear regression,  $r = 0.29$ ,  $p < 10^{-9}$ , Fig. 2.8C). Applying the same analysis to data from PEG showed no relationship between spectral SNR and  $FG_{RG}$  (linear regression,  $r = 0.05$ ,  $p = 0.214$ ).

These results indicate that spectrotemporal interactions between BG and FG stimuli in the auditory periphery tend to produce an overall increase in BG spectral energy. The spectral SNR accounts for some of the FG response reduction in A1 but not PEG. However, even when spectral SNR is 0 in A1, we continue to observe negative RGs (y-intercept, A1:  $-0.27$ , Fig. 2.8C, *left*), indicating that not all the observed effect can be attributed to spectral SNR.

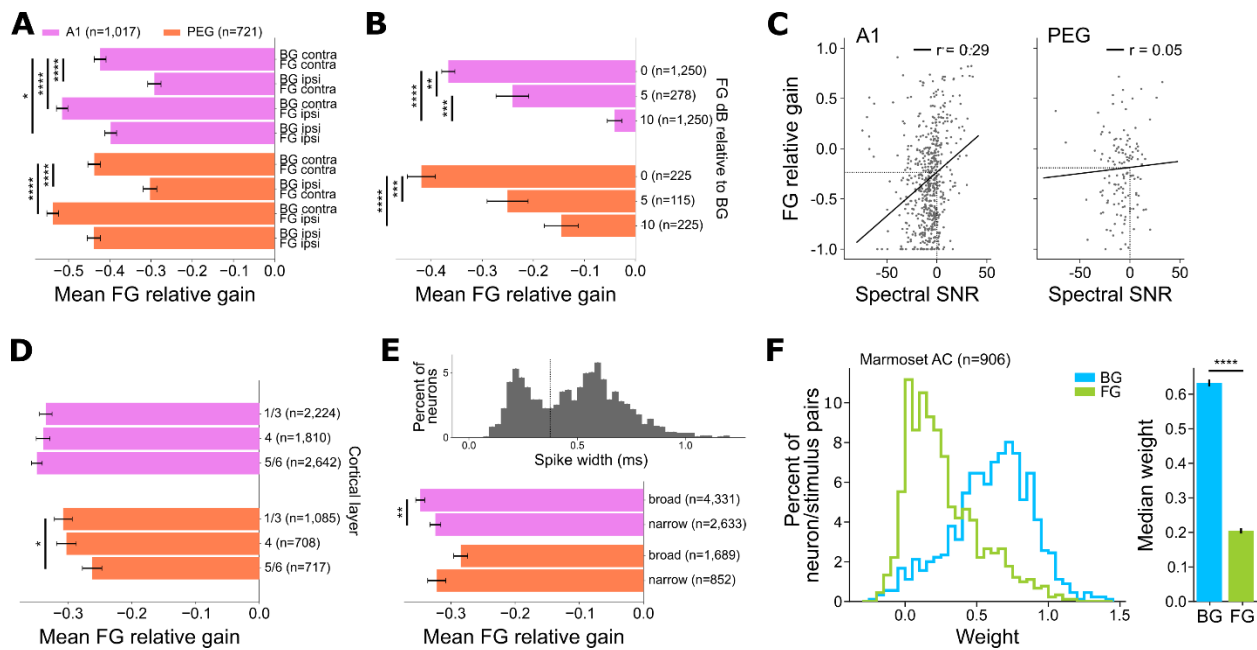
### **Limited influence of cortical depth and cell type on FG-specific response reduction**

Our linear electrode arrays permitted sampling of units across multiple cortical depths. We used current source density analysis (see *Methods*) to classify units according to their approximate cortical layer (layer 1/3, supragranular; layer 4, granular; layer 5/6 infragranular). In general, differences across layers were small. In A1, there were no significant differences between layers. In PEG, FG-specific response reduction in layer 5/6 was modestly decreased relative to layer 1/3 ( $p = 3.46e-3$ , Fig. 2.8D).

In addition, we classified units as broad- or narrow-spiking based on the width of the spike waveform (see *Methods*) (López Espejo and David, 2024). We found a characteristically bimodal distribution of spike widths (Fig. 2.8E, *upper*), with most units in the broad-spiking category (A1: 4,331/6,964, 62.2%, PEG: 1,689/2,563, 65.9%). In A1, but not PEG, FG-specific response reduction was reduced slightly in narrow spiking compared to broad-spiking units (A1:  $p = 9.64e-3$ , PEG:  $p = 0.075$ , Fig. 2.8E, *lower*).

## FG-specific response reduction also occurs in primate AC

Ferrets are evolutionarily distant from humans, and species differences could explain why FG specific reduction has not been previously reported in humans. To determine if the unexpected dominance of BGs in sound mixtures was observed in a more closely related species, we measured the basic BG/FG contrast for single unit activity recorded from AC of the common marmoset (*Callithrix jacchus*). Marmosets are a new world monkey with extensive vocal communication behaviors and AC similar in structure to humans (Bendor and Wang, 2005; de la Mothe et al., 2006; Osmanski et al., 2013; Song et al., 2016; Feng and Wang, 2017). Data from two marmosets recapitulated our observation of strong FG-specific response reduction in AC ( $w_{BG}$ : median =  $0.632 \pm 0.010$ ,  $w_{FG}$ : median =  $0.203 \pm 0.008$ ,  $p < 10^{-9}$ ,  $n = 906$ , Fig. 2.8F).



**Figure 2.8.** FG-specific response reduction across multiple experimental conditions. *A*, Average FG relative gain ( $RG_{FG}$ ) for combinations of contralateral (contra) and ipsilateral (ipsi) BG and FG sounds, grouped by cortical field. (A1:  $n = 1,017$ , PEG:  $n = 721$ , mean  $\pm$  S.E.M, Wilcoxon signed-rank test,  $*p < 0.05$ ,  $****p < 10^{-9}$ ). *B*, Average  $RG_{FG}$  for variable FG sound level (dB), relative to BG, plotted as in *A* ( $***p < 0.001$ ). *N* for each group indicated in the figure. *C*, Scatter plots compare the relationship between spectral signal-to-noise ratio (SNR) at each neuron's best frequency for FG/BG combinations and FG relative gain (A1: linear regression,  $r = 0.29$ ,  $p < 10^{-9}$ , PEG:  $r = 0.05$ ,  $p = 0.214$ ). Dashed lines indicate y-intercept at 0 dB SNR. *D*, Average  $RG_{FG}$  grouped by cortical layer, plotted as in *A* (Mann-Whitney U rank test,  $*p < 0.05$ ). *E*, (*upper*) Histogram of spike width in A1 shows bimodal peaks corresponding to putative inhibitory (narrow width) and excitatory (broad width) neurons. (*lower*) Average  $RG_{FG}$  in for neurons grouped by spike width (Mann-Whitney U rank test,  $**p < 0.01$ ). *F*, (*left*) Histogram of BG and FG weights for all neuron/sound pair combinations recorded in marmoset AC. (*right*) Average FG and BG

weights (median  $\pm$  jackknifed S.E. across neuron/sound pairs,  $n=906$ , Wilcoxon signed-rank test, \*\*\*\* $p < 10^{-9}$ ).

## **Discussion**

We examined single-unit representations of concurrent natural background (BG) and foreground (FG) sounds in ferret auditory cortex (AC). Presenting BG textures concurrently with dynamic FG stimuli resulted in unexpected, selective reduction of FG responses in primary (A1) and secondary AC (PEG). This reduction was dependent on statistical features unique to natural FG sounds. The magnitude of FG reduction was also smaller following prolonged stimulation, consistent with adaptation to an ongoing BG. Similar reduction was observed in marmoset AC, indicating that the phenomenon occurs widely across species. An enhanced, distributed representation of BG stimuli at early stages of auditory processing may be necessary for grouping sound features into perceptual objects (McDermott and Simoncelli, 2011; Shamma et al., 2011; Tye et al., 2024). Once BG features are grouped, they can be subtracted from population-wide activity, producing enhanced representation of behaviorally relevant FG sounds at later processing stages (Mesgarani and Chang, 2012; Kell and McDermott, 2019; Khalighinejad et al., 2019).

### **Representation of natural foreground and background stimuli in auditory cortex**

Studies measuring local field potential and BOLD (fMRI) in human superior temporal gyrus (STG) have reported invariant representations of behaviorally relevant foreground sounds in the presence of natural background noise (Kell and McDermott, 2019; Khalighinejad et al., 2019). Data from A1 has been more limited. One fMRI study found that noise-invariance may not be present at this earlier processing stage (Kell and McDermott, 2017). Studies using stimulus reconstruction methods with single-unit data have argued for enhanced FG representations in A1 (Moore et al., 2013; Rabinowitz et al., 2013; Mesgarani et al., 2014). This work has focused on static, synthetic backgrounds, which may explain discrepancies with the current study. As we find, there is less reduction of foreground responses for stimuli lacking natural spectral and temporal statistics (Fig. 2.7C). It is also important to note that reconstruction of the foreground does not imply that information about background noise is not present in the population response (Christison-Lagay and Cohen, 2014; Malone et al., 2017; Ni et al., 2017). One study of natural background/foreground contrast did report that background stimuli are represented in A1 (Bar-Yosef and Nelken, 2007). Our results show not only that backgrounds are represented, but that they are represented more strongly than concurrent foregrounds in A1 and PEG.

Prior studies using static background noise suggest that foreground invariance emerges by sculpting away inputs in spectro-temporal channels that fall out of band from the foreground signal (Mesgarani et al., 2014; Rabinowitz et al., 2013). However, the spectro-temporal features of natural background sounds

can overlap with those of the foreground. Theories of grouping complex sound features suggest that components of a single source must be identified based on coherent activity across subpopulations of neurons tuned to different features (McDermott and Simoncelli, 2011; Shamma et al., 2011). An overrepresentation of the background may reflect a mixed, distributed code for sound stimuli (Tye et al., 2024). This high-dimensional, overcomplete representation may unmask temporally coherent features (Shamma et al., 2011), allowing for subsequent selection of the foreground object for representation in downstream areas.

### **Natural sound statistics impact background/foreground contrasts**

Our study examined statistical features of natural background and foreground sounds that give rise to differential responses. Sound category was readily distinguished by multiple spectral and temporal properties, and the magnitude of these statistics predicted the degree of background dominance for individual background/foreground pairs (Fig. 2.6). Some effects were consistent with prior work. For example, background sounds tend to have broader bandwidth, which can provide energetic masking of the narrowband foregrounds (McDermott and Oxenham, 2008). Consistent with this idea, our results showed background sounds with broader bandwidth and greater spectral overlap produced larger foreground response reduction, though with a relatively small effect size (Fig. 2.6*I*). More strikingly, responses to foregrounds with greater temporal variance were unexpectedly and consistently more reduced (Fig. 2.6*D*). While high temporal variance of foreground stimuli has been reported, their transient (high variance) features are typically expected to pop out perceptually from continuous noise (Moore et al., 2013; Mesgarani et al., 2014; Kell and McDermott, 2019; Khalighinejad et al., 2019).

Confirming the influence of natural sound statistics on foreground responses, sounds synthesized with increasingly randomized statistics produce successively less of a relative decrease (Fig. 2.7*C*). Notably, foreground responses were less reduced even between fully natural sounds and synthetic sounds that preserved all the measured spectral and temporal properties. This observation indicates that natural sounds contain higher-order properties that further account for background dominated responses.

The statistics investigated here were chosen to capture qualities that categorically distinguished backgrounds and foregrounds (Fig. 2.6*A-F*) and to align with manipulations imposed on synthetic sounds (Fig. 2.7*B*). Other statistics have been proposed for classifying natural stimuli, and they may capture additional differences between background and foreground categories. Measurements of density, sparseness, and burstiness have been used to distinguish sound textures, but remain unexplored when differentiating sound textures from qualitatively distinct foreground-like transients (Zhai et al., 2020;

Mishra et al., 2021). Thus, expanded analysis of sound statistics may provide additional insight into the mechanisms driving the preferential reduction of foreground responses by backgrounds.

Processes shaping the relative response to foreground versus background are likely to also depend on experience and behavior. For the subset of ferret vocalizations, we observed less response reduction than for other foregrounds (Fig. 2.6H), a difference that could not be explained by any of the measured sound statistics. Conspecific vocalizations are likely to have stronger behavioral salience, and prior studies of streaming have described top-down enhancement of behaviorally relevant stimuli (Ding and Simon, 2012; Mesgarani and Chang, 2012; O'Sullivan et al., 2013, 2015; Mesgarani et al., 2014). Alternatively, there may be additional, unmodeled statistical properties of vocalizations that selectively support their relative enhancement in A1. Experiments in trained and behaving animals can determine if behavioral relevance impacts the dominance of background sound responses.

### **Mechanisms shaping the neural representation of concurrent sounds**

Our analysis of the interaction between stimulus statistics and relative gain provides insight into possible mechanisms that shape responses to concurrent sounds. Even the most unstructured synthetic stimuli, lacking natural spectral and temporal modulation, still evoked relatively reduced foreground responses (Fig. 2.7C). These synthetic background sounds maintained a broader power spectrum than foregrounds, which likely activates a wider range of tonotopic channels. Lateral inhibition is a widespread property of cortex (Suga, 1995; Biebel and Langner, 2002; Goense and Feng, 2012; Warren et al., 2013), and broadband stimuli are likely to engage this inhibition to suppress the narrowband foreground response. This hypothesis is bolstered by our results showing reduced foreground responses even when the background did not elicit a significant response on its own (Fig. 2.3C), suggesting that backgrounds evoke subthreshold inhibition that suppresses responses to the concurrent foreground.

A contributing factor to the dominance of the background response could be bottom-up processes in the peripheral auditory system that emphasize spectrotemporal energy of the background relative to foreground. While there was no relationship between responsiveness and metrics of FG reduction (Fig. 2.3B), there was a relationship between spectral signal-to-noise ratio (SNR), which reflects coding in the auditory nerve, and the relative gain of the foreground response in A1 (Fig. 2.8C). However, relative gain remained negative at 0 dB spectral SNR, indicating that the reduced response is not entirely inherited from the periphery. Regardless of its source, the reduced foreground response in A1 and PEG contrasts strongly with the relative enhancement reported in human STG.

Our experiments investigating the dynamics of adaptation to concurrent sounds provide further evidence for the importance of network activity in shaping responses. Studies of speech coding in noise in



human AC have shown that responses adapt to a noisy background over a few hundred milliseconds (Khalighinejad et al., 2019; Mischler et al., 2023), consistent with the single-unit data reported here. Adaptation time following the onset of a background substantially exceeded that of a foreground (Fig. 2.4C, G). Slower adaptation to background sounds is likely a consequence of the background activating a larger portion of the network and therefore requiring longer to reach a steady state than a narrowband foreground.

Dynamics of preferential foreground reduction also differed across the processing hierarchy. Both A1 and PEG showed similar reduction in foreground responses immediately after sound onset, but PEG showed less reduction in the latter half of the stimulus (Fig. 2.5). This relative change may reflect a step toward enhanced foreground representation in downstream areas. Foreground responses in A1 versus PEG also differed in their dependence on sound statistics. Responses to sounds with large spectral correlation were strongly reduced in A1 but less so in PEG (Fig. 2.6I). Neurons in secondary AC are more selective for complex spectral patterns than A1 (Rauschecker et al., 1995; Atiani et al., 2014; Kikuchi et al., 2014; Kline et al., 2021). A neuron tuned to a precise foreground pattern may be less susceptible to interference from a background than an A1 neuron tuned to fewer or less complex features.

### 3. Single-unit representations of background/foreground contrasts in trained, behaving ferrets

In Chapter 2, we answered fundamental questions regarding the auditory streaming of natural sounds at the single-unit level: Responses to natural foreground sounds are unexpectedly and preferentially reduced when presented in a mixture with natural noise. The degree of this preferential reduction of foreground responses by competing background noise could be explained by natural spectral and temporal statistics distinguishing background and foreground categories. Mechanisms producing this response reduction of foreground sounds are complex and cannot be easily explained by the tuning properties of an individual neuron, implicating selectivity at the level of the local network.

In the discussion of these results however, a possible limitation arose in the passive nature of stimulus presentations. Although the stimulus configuration of a background/foreground contrast was specifically selected because of evidence indicating that this contrast induces auditory streaming, ferrets in this study were untrained such that no sounds had any explicit behavioral salience. Thus, there was no precise way to be sure that dynamic foregrounds actually had increased salience to the untrained subject. In the subset of foregrounds containing ferret vocalizations in Chapter 2, the preferential reduction of foreground was still the dominant pattern but to a lesser extent than foreground subsets containing other animal vocalizations or non-vocalizations (Figure 2.6H). These results indicate that the relative reduction of foreground responses by background sounds may indeed depend on experience and behavior, as conspecific vocalizations are likely to hold inherent salience to even untrained animals. Thus, our results were unable to conclude the impact of behavioral salience on relative foreground response reduction.

As a result, Chapter 3 aims to specifically address this gap by adapting the previously presented general stimulus paradigm to a behavioral task to be performed by ferrets which would require successful streaming of natural foregrounds over backgrounds. By recording neural activity in trained ferrets as they perform a spatial streaming task, the impact of experience and behavior on single-unit representations of natural background/foreground mixtures can be examined more closely.

---

#### Single-unit representations of natural foregrounds and backgrounds in an auditory streaming task

Gregory R. Hamersky<sup>1,2</sup>, Jonah D. Stickney<sup>2,3</sup>, Jereme C. Wingert<sup>2,3</sup>, Stephen V. David<sup>2,\*</sup>

<sup>1</sup>Neuroscience Graduate Program, Oregon Health and Science University, Portland, OR 97239, USA

<sup>2</sup>Oregon Hearing Research Center, Oregon Health and Science University, Portland, OR 97239, USA

<sup>3</sup>Behavioral and Systems Neuroscience Graduate Program, Oregon Health and Science University, Portland, OR 97239, USA

## **Acknowledgements**

The results presented in this chapter would not have been possible without the help of Jonah-nah Stickney. I endlessly appreciate his assistance in gathering data. I also must particularly acknowledge his perseverance in the face of the many physical and emotional hurdles presented, both ferret-based and human-based (10%/90%). I will treat him to many games of pinball as an expression of this gratitude.

## **Abstract**

A key challenge of successfully navigating an acoustic world is the ability to understand and attend to behaviorally salient foreground sounds (speech, vocalizations) amidst background noise (environmental noise, mechanical noise). While behavioral salience can be inherent—as in recognizing conspecific communication sounds—it can also be continuously learned and unlearned. Previous studies have shown that high-order areas of human auditory cortex (AC) form enhanced representations of behaviorally relevant foregrounds compared to unimportant noise both pre-attentively as well as via active, top-down processes. Still, evidence has also shown highly encoded background representations in single units. These results, however, have largely relied on statistical categorical distinctions rather than explicit behaviorally motivated relevance. To study the representations of this foreground/background contrast as an active process within a behavioral state, here we recorded single-unit responses in AC of free-moving ferrets during a target-in-noise discrimination task. We trained ferrets to associate and select certain exemplars of a the statistically distinct foreground category as a rewarded target sound over noisy background distractors to determine how training shapes neural responses at early processing stages. Ferrets reliably performed this streaming task under a variety of spatial and difficulty conditions. Neural data from trained ferrets presented with non-task stimuli revealed preferential reduction of foreground responses that was consistent with untrained animals on matched stimuli. We then recorded from a trained, behaving animal and found a relative decrease in the magnitude of preferential foreground response reduction to trained task stimuli, potentially revealing adaption of AC responses to enhance the behaviorally relevant target stimuli. Overall, there were a greater number of noise-invariant responses in trained animals. We hypothesize explanations for this change, but the work presented here is currently limited to a case study due to limited neural data gathered in a second behaving animal. Thus, conclusions with greater statistical power about the extent to which training impacts neural representations of natural targets and distractors will require recordings from additional trained subjects.

## Introduction

As listeners experience the world, they are routinely faced with complex auditory scenes containing temporally and spectrally overlapping sounds. To make sense of this aural hodgepodge, listeners must be able to group cohesive sound features in the time and frequency domains amongst distinct sources, an ability called auditory streaming (Bregman, 1990; Nelken et al., 1999; Carlyon, 2004; Griffiths and Warren, 2004). While these grouping cues have been traditionally identified through psychoacoustic studies probing the boundaries of the statistical relationships that give rise to streaming often using artificial stimuli (Bregman et al., 2000; Bizley and Cohen, 2013), the processes guiding these perceptual groupings are thought to reflect internalized regularities of the natural environment (Młynarski and McDermott, 2019). Thus, auditory systems may be best adapted to accurate streaming through a lifetime of learning those features usually co-occurring within a sound in addition to which sound sources are behaviorally the most relevant.

The importance of behavioral salience in streaming has been shown to be engaged pre-attentively during the perception of salient stimuli in noise (Sussman et al., 2007; Mesgarani et al., 2014; Kell and McDermott, 2019) as well as through active, top-down processes like the selective attention to the voice of a single speaker (O’Sullivan et al., 2015). A useful framework to study natural sound streaming is the background/foreground contrast, as behaviorally relevant foregrounds (e.g., speech, vocalizations) are preferentially perceived over noise-like backgrounds (e.g., waterfalls, fire, machinery) (Bregman, 1990). Indeed, local field potential (LFP) and functional imaging (fMRI) data from human superior temporal gyrus (STG) show that evoked activity in this high-order area of AC preferentially streams foregrounds over backgrounds (Kell and McDermott, 2019; Khalighinejad et al., 2019). In the primary auditory cortex (A1), evidence is mounting that backgrounds are also robustly encoded in single-unit representations, whereby a complete representation of noise may be subtracted from population-wide activity to allow for enhanced representation of foregrounds at later processing stages (Bar-Yosef and Nelken, 2007; Hamersky et al., 2023). The effect of task-specific attention on these single-unit representations of natural background/foreground contrasts in A1 is less known.

Changes in the neural activity of AC have been widely reported after learning an auditory task with pure-tone targets, with neurons showing increased firing rate and reliability in response to tones near the target frequency while suppressing spectrally nearby distractors (Diamond and Weinberger, 1986; Kisley and Gerstein, 2001; Hui et al., 2009; David et al., 2012; Schwartz and David, 2018). It remains unknown whether natural sounds, which produce more distributed activation patterns in AC (Maor et al., 2019), induce similar tuning changes in the AC during learning. Tasks that require the streaming of foregrounds in the presence of natural background noise would be required in animal models amenable to invasive

neurophysiological recordings. Fortunately, streaming-like behaviors have been widely demonstrated in a range of animal models (Itatani and Klump, 2017), including ferrets (Ma et al., 2010).

To measure the effect of behavior and experience on representations of natural background/foreground contrasts, we recorded single-unit activity in primary AC of freely moving ferrets as they performed a two-alternative forced choice task discriminating and locating foreground vocalization targets during concurrent natural background noise. We find reliably strong performance in this task which predictably decreases with increases in difficulty imposed through a variety of spatial configurations and decreased signal-to-noise ratios. We show that while our trained animals have fundamentally similar neural representations where responses to non-task foreground sounds are relatively reduced by natural background noise, trained animals show a higher prominence of noise-invariant neurons to task stimuli. In trained animals, the extent of this increase may depend on task performance, potentially revealing task-related adaptation in the AC to natural sounds.

## **Methods**

### **Surgical Procedures**

All surgical procedures and animal care were performed according to the same protocols, methodologies, and with the same oversight from regulatory bodies as described in Chapter 2.

### **Acoustic Stimuli**

Digital acoustic signals were transformed to analog (National Instruments), amplified (Crown), and delivered through free-field speaker (Manger) placed centrally 80 cm from the front of the arena at 0° elevation and at a 30° azimuth. Stimulation was controlled using custom MATLAB (<https://bitbucket.org/lbhb/baphy>) or Python software (<https://github.com/LBHB/psilbhb>) and all experiments took place inside a custom double-walled sound-isolating chamber (Professional Model, Gresch-Ken).

Auditory stimuli were chosen from a pool of 40 natural sound excerpts, each 2-3 s in length and curated to contain power immediately at onset. Sound excerpts were categorized as two categories, backgrounds (BGs, 20 excerpts) and foregrounds (FGs, 20 excerpts), based on ethological relevance. FGs consisted of dynamic ferret vocalizations (dooks, kit squeaks, fighting) while BGs were natural noise textures (running water, machinery, etc.). Categories could also be distinguished based on simple spectro-temporal statistics described previously (Chapter 2). All sounds were root mean square (RMS) normalized to impose a baseline 0 dB signal-to-noise ratio (SNR) between BG and FG categories, and sound level was calibrated so sounds were presented at a baseline of 55 dB SPL. D

## Two Alternative Forced-Choice Task

Ferrets were trained to perform a spatial two alternative forced-choice task, a target-in-noise detection task requiring locating a target FG sound amidst a noise-like BG. While free-moving, subjects initiate a trial by poking their nose into a port centered between two speakers angled 30° to the left and right in the azimuth, at 0° elevation, and 80 cm from the front of the arena. Two lick spouts were positioned in front of each speaker to either side of the nose poke port. IR detection beams in the nose poke port and each lick spout detect animal responses.

To successfully initiate a trial, an animal must remain in the nose poke port for 0.4 s, at which point a BG sound will begin playing from either or both speakers concurrently to a unilateral presented FG target. Training stimuli were chosen from the same set of 20 BGs and 20 FGs above. An early withdrawal from the port during trial initiation results in an early trial and a brief timeout. A correct response occurs when the animal moves to and nose pokes into the lick spout corresponding with the location of the speaker playing the target FG within 4 s of trial onset, resulting in the delivery of a 0.05 mL 3% sucrose reward. Incorrect trials or trials with no response resulted in a brief timeout where trials could not be reinitiated. Behavioral experiments were performed using customized Python software, `psiexperiment` (<https://github.com/LBHB/psilbhb>).

We tested several variations of target/noise spatial configurations (schematized in Figure 3.1D, in all trials with a presented target a correct trial is defined as a nose poke in the lick spout corresponding with the speaker playing the target FG): (1) *Diotic*: BG was presented from both speakers while a FG was presented from only a single speaker, (2) *Ipsilateral*: BG and FG were both presented from the same speaker, (3) *Contralateral*: BG and FG were presented from opposite speakers. As controls, FG alone trials were presented with the FG in isolation from a single speaker and BG alone trials where the BG plays from a single or from both speakers in the absence of a FG. In BG alone trials, a random lick spout is rewarded with 50% probability. In any of these conditions, the signal-to-noise ratio could be varied such that FG level is attenuated 5, 10, 15, or 20 dB relative to the BG to increase difficulty.

During training, 2-3 sounds per category were selected on a rotating basis from a pool of 20 BGs and 20 FGs. In experiments, all 40 sounds were presented in isolation at each recording site and the 2-3 sounds from each category that evoked the largest average multi-unit response across the site were selected to be used in experiments. Chosen sounds were combinatorically paired to produce unique BG/FG combinations which were presented interleaved amongst the configurations detailed above.

## Neurophysiology

Neurophysiology was performed consistent with practices for preparation of craniotomies and AC mapping using anatomical and tuning properties. In these experiments, recordings were performed by semi-chronically implanting a 960-channel Neuropixel probes (Jun et al., 2017) to permit recordings during free moving behavior and recordings at the same site over multiple recording sessions. Typically, about 150 of the 384 active probe channels spanned AC depth, as determined by current source density analysis (Hamersky et al., 2023). Once inserted, the microdrive was glued down (Flow-It ALC, Pentron) and the craniotomy sealed with a silicon polymer (Kwik-Cast, World Precision Instruments). A custom 3D printed enclosure was installed around the microdrive to protect the implant. Data were amplified, digitized, sorted, and manually curated to analyze only single units as in Chapter 2.

Neurophysiology recordings were performed in a freely moving, behaving state. As a comparison, task stimuli were presented to passive, head-fixed ferrets during the same recording session. Sessions typically lasted 4-6 h. In each recording session, video and LFP monitored the animal's state to monitor the animal's state. After 2-3 days of recording, the probe was explanted.

## Inclusion Criteria

We measured evoked activity with the peri-stimulus time histogram (PSTH) response averaged across sound repetitions at 100 Hz sampling. All passive recordings had 10 sound repetitions and when analyzing behavior trials needed at least 5 completed repetitions per sound to be included in analysis. As in Chapter 2, we placed numerous restrictions on stimulus/neuron pairs to ensure that we only include sound-responsive units in analysis.

We calculated a signal-to-noise ratio for each stimulus/neuron pair using the ratio of the PSTH response to the standard deviation of the response across repetitions (Fritz et al., 2003). For experiments shown in Figure 3.2A, our inclusion criteria matched those of Chapter 2, requiring an  $\text{SNR} \geq 0.12$  to be considered sound responsive. In behavioral analysis (Fig. 3.2D), we found overall lower SNR amongst units, likely because of fewer sound repetitions, so we expanded our cutoff for sound responsiveness to  $\text{SNR} \geq 0.08$ . We considered a neuron/stimulus pair to be responsive when responses to both BG and FG in isolation exceeded threshold (628/2,618, 24%). We similarly only included responsive units where the linear model fit well ( $r \geq 0.4$ , 99/628, 16%). In most behavioral analyses (such as Fig. 3.2D, *right*) we required all SNR and fit conditions to be met in both the behaving and passive conditions. These strict criteria severely limited our included neuron/stimulus pairs.

We discuss the presence of motor neurons in our recordings (Fig. 3.3), which showed robust responses particularly in the FG and BG+FG behavior conditions but little to no response during the passive, head-fixed recording of the same sounds. To best isolate motor neurons and remove them from our weight analysis, we computed a normalized difference between responsiveness of the behaving and passive PSTHs during the 1 s fit window:

$$Relative\ difference = \frac{\mu_{behaving} - \mu_{passive}}{\sqrt{\sigma_{behaving} + \sigma_{passive}}}$$

This metric quantified how much greater the response in the behavior condition was relative to the passive. As a result, we could ascribe a cutoff of 1.1 at which we considered units that dissimilar between conditions to be motor neurons (13/99, 13%).

## Statistical Analysis

In pairwise statistical tests (Figs. 3.2A, 3.2D) we performed a Wilcoxon signed-rank test with significance determined at the  $\alpha = 0.05$  level. We note the number of neuron/sound pairs in figures and figure legends along with exact p-values. In unpaired comparisons (Fig. 3.2D) we used a Mann-Whitney U rank test with significance at the  $\alpha = 0.05$  level.

## Results

### Ferrets can learn a streaming behavioral task with natural sounds

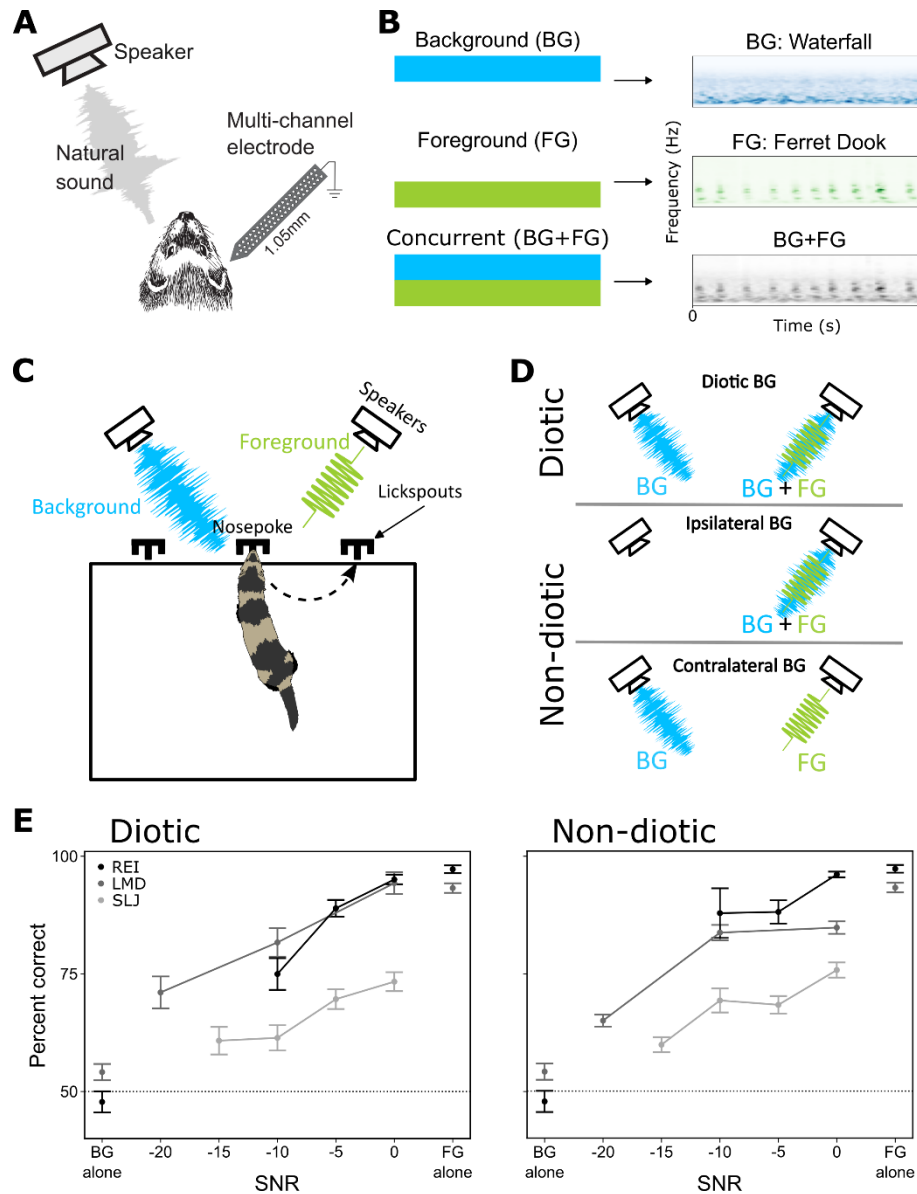
To explore how neurons in the auditory cortex (AC) integrate information about behaviorally relevant sounds in the presence of noise during behavior, we trained ferrets to perform a two-alternative forced choice (2AFC) task where a target sound must be spatially located in the presence of background noise. We would compare responses to targets paired with distractors in back-to-back stimulus-matched passive (Fig. 3.1A) and behaving (Fig. 3.1C) trial blocks to directly determine the effect of task engagement on target representations.

Target and noise stimuli were drawn from natural sounds comprising two ethological categories: background textures (BGs) and ferret vocalizations (FGs). A BG/FG contrast stimulus configuration was chosen because it has been well established as producing streaming (Rabinowitz et al., 2013; Mesgarani et al., 2014; Kell and McDermott, 2019; Khalighinejad et al., 2019). Ferret vocalizations were chosen as their own target category due to the inherent behavioral salience of conspecific calls, with diverse exemplars from a range of behaviors comprising this category.



During training on a two-alternative forced choice (2AFC) task, 2-3 BG and FG exemplars (2-3 s duration) each were chosen on a rotating basis to be presented in isolation or simultaneously (Fig. 3.1B). Trials are initiated when a ferret inserts their nose into a port located centrally between two speakers to break an IR beam (Fig. 3.1C, for complete details, see *Methods*). In general, a test trial consists of a target (FG) played from a single speaker and a distractor (BG) played from one or both speakers. A correct trial entails the subject moving to and nose-poking into the lick spout that corresponds to the speaker playing the target FG, at which point a sucrose reward is delivered. To modify difficulty, BGs could be oriented in one of three configurations: (1) diotic – BG played from both speakers while FG plays from one, (2) ipsilateral – BG played from the same speaker as the FG, and (3) contralateral – BG played from the opposite speaker as the FG (Fig. 3.1D). We pool and collectively refer to ipsilateral and contralateral trials as non-diotic. Sound mixtures were presented with the FG attenuated at various levels relative to the BG to increase difficulty as SNR decreases. Some trials also consisted of control trials where the FG played in the absence of a BG or where the BG played in the absence of a FG. In the latter case, either lick spout was rewarded with a 50% probability.

Animals (n=3) were able to perform the task well above chance level in both diotic (0 dB, REI:  $95.0 \pm 1.0\%$ , LMD:  $94.3 \pm 2.3\%$ , SLJ:  $73.3 \pm 2.0\%$ ) and non-diotic (0 dB, REI:  $96.0 \pm 0.6\%$ , LMD:  $84.7 \pm 1.3\%$ , SLJ:  $75.7 \pm 1.6\%$ ) conditions, where we observed performance decrease steadily with increasing SNR difficulty (Fig. 3.1E). As expected, performance on FG alone trials was near perfect performance in both animals tried in this condition (REI:  $97.2 \pm 0.8\%$ , LMD:  $93.2 \pm 1.0\%$ ) while BG alone trials without a target were at near chance performance (REI:  $47.8 \pm 2.2\%$ , LMD:  $54.1 \pm 1.7\%$ ). Having attained reliable behavioral performance on 2/3 animals (REI and LMD) on this streaming task, we next recorded from AC of animals during task performance to determine how task engagement may affect representations of the BG/FG contrast.



**Figure 3.1.** Behavioral task experimental setup and performance results. **A**, In the passive complement to behavioral recordings, head-fixed ferrets were presented natural sound stimuli from a free-field speaker  $30^\circ$  contralateral to the recording hemisphere. Neuropixel arrays recorded single-unit activity from the primary (A1) field of AC. **B**, During recording, natural sound excerpts from two distinct, ethological categories—backgrounds (BGs) and foregrounds (FGs)—were presented in isolation (blue and green spectrograms, respectively) and concurrently (black spectrogram). **C**, Schematic of two-alternative forced choice task (2AFC). In all variations, subjects initiated a trial with an extended nose poke to the center port. A trial consists of a presentation of a target (FG) and distractor (BG), and a correct trial is when the subject moves to the lickspout corresponding with the speaker playing the FG (shown as the dotted arrow). **D**, Different spatial variants of the 2AFC, where BG is presented in different spatial orientations relative to the FG. **E**, Results from behavioral training of animals ( $n=3$ ) on the 2AFC task in the spatial configurations outlined in **D** (reported as percent correct  $\pm$  SEM). \*Panels **C**, **D** modified with permission from JCW.

## Responses of trained animals to non-task natural foreground/background pairs are similar to untrained animals

We next recorded single-unit activity in 2/3 trained ferrets (LMD and REI). We first compared interactions in behaviorally trained ferrets to our untrained ferrets using a similar stimulus configuration that showed unexpectedly dominant relative reduction of natural FG sounds reported in Chapter 2. Here, we played a standardized set of the same 4 BGs and 4 FGs (paired combinatorically) at each recording site. We used the same inclusion criteria described in Chapter 2 and fit the same linear model to describe the combined BG+FG response as a linear weighted sum of the constituent BG and FG responses in isolation (diagrammed in Fig. 2.2A):

$$R_{\text{BG+FG}}(t) = w_{\text{BG}}R_{\text{BG}}(t) + w_{\text{FG}}R_{\text{FG}}(t)$$

In brief, individual BG or FG weights  $< 1$  indicated a relative reduction of the individual response in the BG+FG response. We then compared  $w_{\text{BG}}$  and  $w_{\text{FG}}$  for each neuron and stimulus pair, or instance, tested. Results in our two new subjects were indistinguishable from previous animals ( $n=5$ ), showing significantly higher BG weights compared to FG weights ( $p < 10^{-9}$ , Fig. 3.2A). Recordings in this stimulus configuration served as a control that our trained animals recapitulated the unexpected result of preferentially reduced FG responses previously observed on stimuli not explicitly trained in the task.

Next, we recorded single-unit activity during the 2AFC task and during passive listening to task stimuli so we could compare the effect of attention on response interactions. At each recording site, we selected 2-3 BG and FG exemplars each from a larger set of 20 BGs and 20 FGs (2-3 s duration). Each individual exemplar was first presented in isolation to determine site responsivity to each sound, and we chose the sounds that evoked the largest average multi-unit response to be presented simultaneously (BG+FG) and in isolation (BG, FG) during experiments. Behavioral performance during recordings was overall consistent with training performance at 0 dB (REI:  $95.7 \pm 1.0\%$ , LMD:  $77.8 \pm 2.3\%$ ), FG alone (REI:  $99.5 \pm 0.5\%$ , LMD:  $91.2 \pm 1.9\%$ ), and BG alone (REI:  $50.7 \pm 2.6\%$ , LMD:  $52.7 \pm 2.9\%$ ) trials (Fig. 3.2B). Of note, because of experimental limitations, the 2AFC task during recording was largely limited to the 0dB, non-diotic condition, which will be the subset of conditions from which the data for the forthcoming analyses were drawn.

## Behavioral performance may directly affect extent of FG response reduction

Response interactions of target FGs and BG distractors were diverse. In many neurons, we observed the profound reduction (Fig 3.3C, *left*) of FG responses reported previously (Fig 2.2) even in the absence of a robust BG response (Fig. 2.3), indicative of network level inhibition by BGs. Other responses were less typical in the context of our previous work. For example, whereas a negligible number of neurons

showing BG-invariance were previously observed, here we note many instances of neurons where responses to the FG and BG+FG were near identical, a  $w_{FG}$  of  $\sim 1$  (Fig. 3.3C, *right*). This could reflect adaptation in the AC to enhance representations of task stimuli (Diamond and Weinberger, 1986; Hui et al., 2009; David et al., 2012).

Group data of fit weights was consistent with this qualitative observation of a relatively increased number of recorded noise-invariant neurons robustly responsive to task foregrounds in trained animals. We compared only weights from instances where reliable responses were evoked to both individual BG and FG sounds in during both active and passive recording blocks (LMD:  $n=341/1,502$ , 23%, REI:  $n=287/1,116$ , 26%, see *Methods*) as well as instances with a good model fit ( $r \geq 0.4$ , LMD: 23%, REI: 7%). To describe the relative contribution of FG versus BG to the BG+FG response, we combine  $w_{BG}$  and  $w_{FG}$  into a single metric, FG relative gain ( $RG_{FG}$ ):

$$RG_{FG} = \frac{w_{FG} - w_{BG}}{|w_{FG}| + |w_{BG}|}$$

With this metric,  $RG_{FG}$  values  $< 0$  indicate FG-specific response reduction, where the FG is relatively more reduced than the BG.  $RG_{FG}$  values  $> 0$  indicate enhanced FG responses. We report uniform reduction of FG responses between active and passive recordings in both animals (LMD:  $n = 65$ , mean active:  $-0.12 \pm 0.07$ , passive:  $-0.13 \pm 0.06$ ,  $p = 0.82$ , Wilcoxon signed-rank test; REI:  $n = 21$ , mean active:  $0.25 \pm 0.14$ , passive:  $0.27 \pm 0.13$ ,  $p = 0.96$ , Fig. 3.2D, *right*). In comparison to previous results, mean  $RG_{FG}$  is higher in both animals under similar SNR conditions (Fig. 2.2), indicative of less FG-specific response reduction, possibly reflecting an effect of behavioral training.

We report unpaired comparisons between the untrained, naïve condition (Fig. 3.2A), with passive and behaving results (Fig. 3.2D, *left*). Neurons in this comparison did not have to meet our inclusions criteria across conditions. We observed several trends in the results consistent between animals. Notably, while neither subject shows a difference between passive and behaving, both do show a substantial relative increase in  $RG_{FG}$  between conditions with task stimuli and the untrained condition with non-task FGs. As a result, both subjects may be showing a similar, task-dependent relative increase in  $RG_{FG}$  between non-task and task stimuli, despite having different non-task baselines.

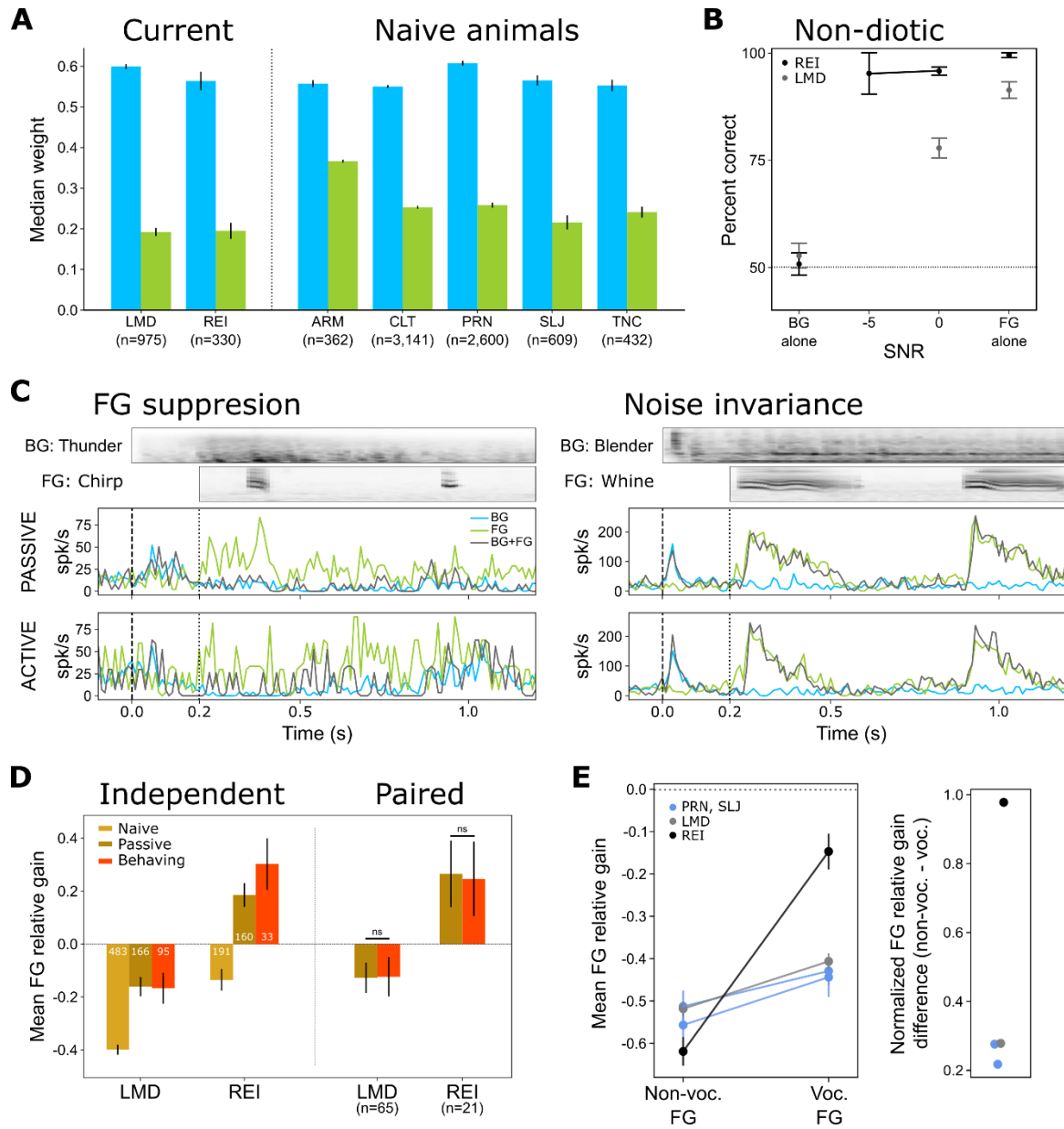
While this difference in baseline could be attributable to different effects of behavioral training or more broad generalization of task and non-task vocalizations in one animal, it is also possible that this divergence between LMD and REI is simply noise because of a very low sample size in REI (neurons/sound pairs  $n = 21$ , recording sites  $n = 2$ ). This restricts our ability to draw additional conclusions comparing the animals currently.

To explore this change in baseline further, we revisited our previous analysis in untrained animals from Chapter 2 that showed the vocalization sub-category of FG sounds with a smaller degree of FG-specific response reduction compared to the non-vocalization FG sub-category (Fig 2.6H), which we hypothesized to be a result of inherent behavioral relevance of conspecific vocalizations. To investigate if this trend was consistent in trained animals, we similarly sub-classified FGs from the standardized set of non-task stimuli to determine if LMD and REI differed in their representations of these sub-categories in this stimulus configuration. Indeed, mean  $RG_{FG}$  of both the non-vocalization and vocalization sub-categories in LMD was nearly identical to that of the untrained animals (LMD: non-voc.  $-0.52 \pm 0.02$ , voc.  $-0.41 \pm 0.02$ , Untrained A: non-voc.  $-0.56 \pm 0.03$ , voc.  $-0.44 \pm 0.05$ , Untrained B: non-voc.  $-0.51 \pm 0.04$ , voc.  $-0.42 \pm 0.03$ , Fig. 3.2E, *left*). Only 2/5 untrained animals contained a large enough sample ( $n \geq 90$ ) of the sounds from the standardized set played to LMD and REI for meaningful comparison. Meanwhile, while the non-vocalization sub-category in REI showed a similar  $RG_{FG}$  ( $-0.62 \pm 0.03$ ), FG-specific response reduction of the vocalization sub-category was much less ( $-0.15 \pm 0.04$ ). The relative  $RG_{FG}$  change between sub-categories was normalized to account for variations in magnitude of FG response reduction amongst different subjects:

$$Relative\ difference = \frac{\mu_{non-voc.} - \mu_{voc.}}{\sqrt{\sigma_{non-voc.} + \sigma_{voc.}}}$$

This normalized difference can be seen as moderate and equivalent between LMD and naïve animals (LMD: 0.28, Untrained: 0.28 and 0.22) while REI showed a drastic change in FG response reduction between sub-categories (REI: 0.98, Fig. 3.2E, *right*).

Together, despite lacking statistical strength in one animal, we observe similar trends of relative FG enhancement between non-task and task stimuli in both animals (Fig. 3.2D, *left*). Additionally, both animals show no difference between passive and behaving trial blocks, suggesting that adaptation to FG targets may not be a rapidly reversible change. These results suggest that natural sounds induce changes in the tuning properties of AC neurons much like how has been demonstrated to pure tone stimuli.

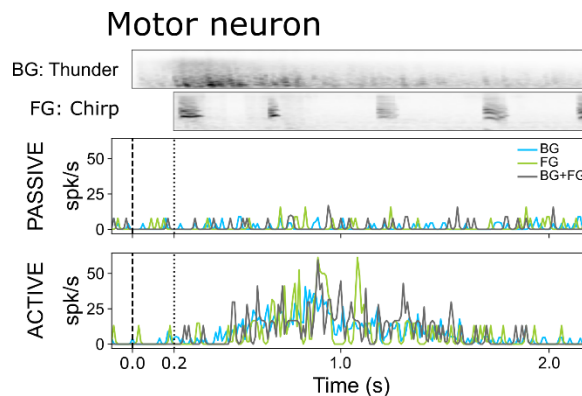


**Figure 3.2.** Response interactions during recording from trained animals. **A**, Comparison of  $w_{BG}$  and  $w_{FG}$  during presentation of non-task BG and FG stimuli in a configuration consistent with previous, untrained subjects. In trained ferrets LMD and REI, median  $w_{FG}$  was significantly lower than  $w_{BG}$  with magnitudes consistent with that of untrained animals. Bars at left compare median weights ( $\pm$  jackknifed S.E. across neuron/sound pairs, Wilcoxon signed-rank test,  $p < 10^{-9}$ ). **B**, Behavioral performance from recording days in the non-diotic spatial configuration. Performance at the 0 dB test condition in both subjects falls in between chance performance at BG alone trials and near perfect performance in FG alone trials. **C**, Example PSTH responses in the behaving and passive complement from two units. BG and FG responses are shown in blue and green, respectively, and the actual BG+FG response is shown in black. One example (*left*) shows FG response reduction while the other (*right*) shows noise-invariance. **D**, Summary of weight analysis in trained animals, expressed in mean  $RG_{FG}$ . On the left, mean  $RG_{FG}$  is shown for each subject from the (1) stimulus naïve, untrained configuration shown in **A**, (2) passive recording blocks and their (3) behavioral complement. Here, each condition must independently meet neuron/stimulus pair inclusion

criteria to maximize sample size. Meanwhile, at right, mean  $RG_{FG}$  is shown only in neuron/stimulus pairs where both the behaving and passive condition met criteria, allowing for paired comparison, which showed no significant differences (Wilcoxon signed-rank test). *E*, (*left*) Mean  $RG_{FG}$  when the results shown in *A* are split according to FG sub-category: non-vocalization or vocalization. Untrained animals are shown in blue while trained animals LMD and REI are shown in grey and black, respectively. (*right*) Relative difference between  $RG_{FG}$  for the two vocalization sub-categories. Colors as in *left*.

### Motor neurons in free-moving recordings

There is growing evidence supporting the presence of neurons encoding motor features like movement and place in the sensory cortices (Holey and Schneider, 2023; Mertens et al., 2023; Mimica et al., 2023). In our analyses comparing responses between behavior and passive listening, we noticed numerous neurons who lacked activity in the passive condition but had robust firing during behavior that were fairly uniform regardless of sound stimuli—likely motor neurons (Fig. 3.3). We used this general distinction between passive PSTH and behaving PSTH to roughly categorize motor neurons in our recordings (see *Methods*), finding a that 7.5% (156/2,076) of recordings single-units met our criteria. The implications of their presence is beyond the scope of this chapter but will be discussed in Chapter 4.



**Figure 3.3.** Example PSTH of a neuron that responds to animal movement in the active condition but shows no response during passive, head-fixed listening. BG and FG responses are shown in blue and green, respectively, and the actual BG+FG response is shown in black.

### Discussion

We explored the effects of behavioral training and experience on single-unit representations of concurrent natural foreground (FG) targets paired with natural background (BG) distractors in ferret auditory cortex (AC). Ferrets learned a novel two alternative forced-choice (2AFC) task that required streaming to successfully identify the target sound amidst a noisy distractor. Trained animals showed comparable preferential reduction of non-task FG responses in a BG+FG mixture during passive listening. Both also showed a relatively degree of FG response reduction to task stimuli compared to their

non-task baseline. Adaptation that enhances neural tuning to task stimuli has been widely reported for pure tones to enhance behavioral performance (Diamond and Weinberger, 1986; Kisley and Gerstein, 2001; Hui et al., 2009) and increase neural discriminability (Heller et al., 2023). Though this adaptation has been shown to be rapid and reversible in a state-dependent manner to pure tone targets (Fritz et al., 2003), here we report the relatively enhanced representation of natural FG target sounds to be uniform between behavior and passive states. This result suggests a difference in the patterns and mechanisms of plasticity in AC neurons to more complex natural sounds.

### **An animal model of an auditory streaming task with natural target and distractor sounds**

Although many studies have demonstrated the ability of animals to perform tasks that require auditory streaming (Itatani and Klump, 2017), few or no paradigms for streaming have been established using natural targets and distractors to create a more ethologically relevant behavior. Our 2AFC task is also unique because rather than training on a single exemplar of a foreground to be used as a target, we require animals to generalize exemplars across a target category, ferret vocalizations. This category is comprised of sounds from across the ferret's repertoire of vocalizations—playful dooks, kits whining, fighting. As a result, unlike past streaming tasks where a pure tone of differing frequencies must be detected, no single spectrotemporal property binds the category of ferret vocalization (Fig. 2.6*A, C, E*), necessitating more abstract grouping.

Attributing the ability of ferrets to generalize an abstract category like ferret vocalizations must be done with caution, however. Because we are using a BG/FG distinction which we have demonstrated to be categorically distinct in spectral, temporal, and cochlear activation patterns (spectral correlation, temporal variance, bandwidth, Fig. 2.6), it is possible ferrets are relying on the increased temporal variance of foregrounds in general, for example, compared to backgrounds rather than the ethologically defined category. To test the strategies employed by animals when distinguishing between target and distractor, further studies might include catch trials where a ferret vocalization target is paired with a non-vocalization foreground as a distractor with similar spectrotemporal statistics. Alternatively, trials containing a background distractor and non-vocalization catch that is statistically like a target foreground could inform us whether there is a bias towards statistically foreground-like sounds. These experiments could help us ask more specific questions about natural sound grouping in ferrets that are beyond our present scope.

Regardless, we show that across a range of spatial configurations and challenging SNRs, behavior in trained ferrets indicates a robust ability to stream natural foreground targets in the presence backgrounds distractors with similarly natural statistics (Fig. 3.1*E*). With this level of task performance, we can



confirm that animals are perceptually streaming FGs during recordings with a certainty we could not during recordings in Chapter 2.

### **Effects of training on the single-unit representations of concurrent backgrounds and foregrounds**

Our study sought to identify how neural representations of FGs are affected by attention in streaming. Receptive fields of neurons in primary AC (A1) readily adapt to enhance tuning around the frequency of pure tones with a conditioned reward (Hui et al., 2009). We report relatively enhanced (increased  $FG_{RG}$ ) responses to task FGs versus non-task FGs (Fig. 3.2D, *left*), confirming that A1 neurons also enhance tuning to natural sounds because of training, a result possibly reflected in the appearance of completely noise-invariant PSTHs (Fig. 3.2C, *right*).

The mechanisms surrounding this enhancement in natural sounds remain less clear. Pure tones have very focal activation patterns, causing similar focal enhancements in neural tuning reflected in STRFs. Natural sounds contain complex spectral content to produce distributed activation patterns in AC (Maor et al., 2019) which may require higher dimensional representations than a linear model can represent. Neurons in AC produce higher-dimensional representations of complex sounds to encode their increasingly complex spectrotemporal patterns (Rauschecker et al., 1995; Kikuchi et al., 2014), including neurons tuned so specially they respond only to categories of sounds like conspecific vocalizations (Montes-Lourido et al., 2021). As a result, it is likely that enhanced tuning to natural FG targets may not be as simple as that of pure tones. As mentioned above, although FG exemplars used as targets come from the same vocalization category, they can be selected from several distinct vocalizations and therefore have no simple, single-dimensional spectrotemporal feature that may be enhanced to improve performance. As a result, the adaptations necessary to enhance representations of the diversity of spectrotemporal characteristics across the vocalization category are likely complex. Determining the exact training-induced changes in AC neurons may require more complex, non-linear models to describe the computations of higher-order neurons (Keshishian et al., 2020).

Relatively complex tuning changes may be why we see no evidence that enhancements in FG representation are limited to the behaving state and are not readily reversible. The STRF changes observed in trained pure tones are limited to the behavioral state and quickly revert to their original state following behavior (Fritz et al., 2003). From this, we would expect that mean  $RG_{FG}$  would be higher during behavior compared to the passive complement, but both animals are indistinguishable between conditions (Fig. 3.2D, *right*). In Chapter 2 we proposed that the preferential reduction of foreground responses by backgrounds in untrained animals may be driven by network activity shaping responses, such as lateral inhibition (Suga, 1995; Goense and Feng, 2012; Warren et al., 2013). Backgrounds have an

overall broader and more uniform power spectrum than foregrounds which is likely to engage more of the network to possibly suppress the responses of comparably narrowband foregrounds. To enhance responses of natural foreground targets then, it is possible that the changes are not only more complex but might also affect the wider network, which could be changes less rapidly reversed compared to the focal, state-dependent STRF changes induced by pure tone training.

### **Evidence of behavioral category generalization**

Our analysis and comparison of background and foreground weights to non-task stimuli in our trained animals and untrained animals showed fundamentally no difference in amounts of foreground response reduction (Fig. 3.2A). This was reassuring, validating our results from Chapter 2 and confirming that changes we observe in task stimuli may be associated with training. Although we saw relatively increased weighting of foreground responses between non-task stimuli and task stimuli in both animals (Fig. 3.2D, *left*), a particularly striking result was the drastic difference in baseline foreground response reduction of non-task sounds between REI and other animals. In the non-vocalization sub-category, REI showed equivalent reduction to LMD and untrained ferrets. However, the vocalization sub-category showed substantially less foreground response reduction than any other animal (Fig. 3.2E), even though these vocalizations were not trained. A change in the representation of untrained exemplars of the target category may reflect broad categorical generalization by this animal.

## 4. Conclusions and future directions

Studies of auditory streaming have indicated that noise-robust perception may arise in auditory cortex (AC) by preferentially emphasizing and constructing invariant representations of behaviorally salient sound features (Rabinowitz et al., 2013; Mesgarani et al., 2014; Khalighinejad et al., 2019). While these studies have focused on what information about relevant sounds can be extracted from sound mixtures, it remains unexplored how these sounds and what information, if any, of the background noise may be represented at the single-unit level in primary AC (A1).

The work presented in this dissertation builds on these studies by directly asking this question to investigate how representations of two statistically and ethologically distinct natural sound categories differ when presented concurrently. In Chapter 2, I showed static background textures in primary and secondary AC dominate responses to dynamic, transient foreground sounds. This result was unexpected given the literature suggesting ubiquitous noise-invariant representations but suggests that it may be necessary to thoroughly encode the noise so it can be subtracted at later processing stages. These results are particularly compelling because the relative reduction of foreground responses was not entirely dependent on neural tuning but was shown to be highly dependent on the natural statistics unique to foreground and background categories.

Then, in Chapter 3 I applied this analysis to an auditory streaming task to determine the effect of behavior and experience on single-unit representations of this background/foreground natural sound contrast when foregrounds were given explicit behavioral salience. I present results showing ferrets can perform a task that requires auditory streaming uniquely using both natural targets and distractors. Further, trained subjects showed a smaller extent of response reduction for natural foreground targets by distractors when compared to non-task foregrounds or untrained animals. This result suggested that tuning in the AC can also adapt to enhance representations of trained stimuli with natural, complex spectrotemporal statistics. The interpretation of this data is currently limited, however, due to low statistical power from a second animal and must be treated as a preliminary case study.

Overall, the work presented uniquely studies single-unit representations of overlapping natural sounds to reveal new insights about how well-studied properties of the AC in streaming may be applied to sounds with natural spectrotemporal statistics. Throughout this dissertation, I've emphasized the surprising nature of our main result that the responses to natural foreground sounds are relatively and preferentially reduced when presented concurrently to natural background noise. At the heart of this surprise was perceptual evidence that would suggest—because auditory streaming allows the perception of salient sounds in a chorus of noise—that foreground sounds also have enhanced neural representations through auditory

cortex to lead to this noise-robust perception. What implications, then, does the result of foreground-specific response reduction have and how might an overrepresentation of background noise at earlier stages of the auditory hierarchy ultimately facilitate noise-invariant perception?

## **4.1 Natural spectrotemporal statistics play a crucial role in the interactions of foreground and background sounds**

In my Introduction, I gave a tongue-in-cheek conclusion to the section overviewing the broad differences between synthetic and natural sound stimuli (1.2.2), imbuing the latter with a *je ne sais quoi* that the former could never attain. Indeed, the literature has supported this notion not just through broad differences comparing simple, synthetic sounds with complex, natural sounds, but also by directly showing that even synthetic sounds extensively curated to model natural spectrotemporal relationships do not match their natural counterparts perceptually (Młynarski and McDermott, 2019) or through elicited neural response (Norman-Haignere and McDermott, 2018). Taken with the results I have presented, what seemed like a flourish of writing has become a crucial part of the work in this dissertation as it relates to understanding how the AC represents mixtures of overlapping natural sounds.

In Chapter 2, we calculated statistics of natural sounds that describe their cochlear activation patterns and their spectral and temporal relationships. We found ethologically distinct foreground and background categories to differ by these statistical metrics—foregrounds activated more narrow regions of the cochlea, were more transient in time, and contained greater correlations between frequencies than backgrounds. These statistical differences, independent of neural tuning, predicted the reduction of responses to the statistically more foreground-like sounds by statistically more background-like sounds (Fig. 2.6*B, D, F*). More directly, when we synthesized sounds to match different combinations of these three natural statistics, we found that even the most “natural” synthetic sound showed different background/foreground interactions and matching progressively fewer statistics largely erased the categorical distinction between foregrounds and backgrounds (Fig. 2.7*B*) and dramatically reduced foreground-specific response reduction (Fig. 2.7*C*).

Generally, these results validate our unique paradigm studying auditory streaming using natural foregrounds and natural background noise. Previous studies frequently use natural foregrounds because perceptually those are the important ones, but they often also use synthetic noise to simplify the categorical contrast (Mesgarani et al., 2014). This stimulus selection frames the background/foreground contrast with the implicit bias of our perceptions that foregrounds are the important ones and, lo and

behold, the story becomes a reflection of that with evidence showing the foreground can be decoded from the mixtures thus precluding us from caring about what happened to the background.

I argue here that while simplifying noise is convenient, it may not answer the question I am after. A lifetime of experience shapes our acquisition of natural sound statistics and their relationships which in turn affects our internalized template of our natural environment (Młynarski and McDermott, 2019). Therefore, to study how an ethological acoustic scene is represented, it is important to use a naturalistic scene. By doing so, we can allow the brain to dictate which aspects of sounds are most important for normal perception, an especially important point when considering a task like streaming that constantly requires the parsing of interactions of natural sounds. Here, by presenting the brain with equally naturalistic backgrounds and foregrounds, we found an unexpected result showing that neurons do in fact robustly represent noise (Chapter 2) in a way that may be changed based on experience (Chapter 3).

In the same way simplifying natural sounds to synthetic sounds is convenient, I too am guilty of employing a simplification: the background/foreground contrast. Though well established to produce streaming (Rabinowitz et al., 2013; Mesgarani et al., 2014; Kell and McDermott, 2019; Khalighinejad et al., 2019) and quantitatively distinct through several statistical metrics (Singh and Theunissen, 2003; Kell and McDermott, 2019; Attias and Schreiner, 1997), in a more natural context the distinction between background and foreground is more arbitrary—maybe you do actually want to listen to the serene yet statistically static sounds of a flowing waterfall while ignoring the dynamic, transient voices of the tourists crowding Multnomah Falls. Having demonstrated sound statistics to be a large contributing factor of the interactions between our natural backgrounds and foregrounds, an extension of this work explores these statistics on a continuum, unconstrained by category.

To this end, I have laid the groundwork for an experiment that addresses this question directly, focusing on the three statistics described in Chapter 2: bandwidth, temporal stationarity, and spectral correlation. I compiled a collection of ~600 natural sound excerpts comprised of sounds that would neatly fit into background/foreground as well as more categorically ambiguous cases such as coins rattling or ensemble music. I calculated the three metrics for each sound, tiling the 3d space defined by these statistics. In a passive-listening experiment, I can now identify a singular axis through this space and pair sounds along this axis in our most basic concurrent stimulus configuration. For example, if we select a Sound A that is more narrowband, more spectrally coherent, and temporally transient (hallmark traits of a foreground) and pair it with Sound B that is broadband, less spectrally coherent, and temporally stationary (hallmark traits of a background) we can expect based on our previous results to see a preferential reduction of Sound A. But, free of categorical limitations, we can now replace Sound B with Sound C that falls as a direct statistical intermediate between A and B and ask what the interaction between A and C

will now be? Will responses to Sound A remain reduced albeit to a much smaller extent because of the more moderate statistics of Sound C? Anecdotal evidence from experiments in Chapter 2 where a relatively narrowband and temporally transient background, rocks tumbling down a hill, showed some of the smallest relative gain values suggests this would be the case. Further, what happens if we pair two sounds that are statistically equivalent in bandwidth and spectral correlation but differ in temporal properties? In this configuration our exploration of how spectrotemporal properties of natural sounds affect their interactions can be more granular and the relationships more controlled, allowing us to delve into even more naturalistic scenarios and more interesting, dynamic interactions.

## 4.2 The role of binaural cues in natural sound representations

Our analyses in Chapter 2 focused almost entirely on natural background/foreground stimuli presented from the same speaker positioned contralateral relative to the recording hemisphere of the head-fixed ferret. Experimentally, this was important to control the relative intensity of background and foreground sounds as we studied their interactions. We did, however, include the modification where backgrounds and foregrounds were presented in different spatial combinations from the contralateral speaker as well as another opposite speaker ipsilateral to the recording hemisphere.

Although spatial location plays a smaller role in streaming to supplement other monaural streaming cues (Shinn-Cunningham, 2005), we included this modification to explore if spatial location would affect the preferential response reduction of foreground sounds. Our results were consistent with this as well as with prior results indicating preferential encoding of contralateral stimuli by AC (King and Middlebrooks, 2010)—the degree of foreground response reduction was smaller when the foreground was contralateral and relatively louder than an ipsilateral background and increased in the opposite spatial configuration (Fig. 2.8A).

The role of binaural cues in streaming had a much more prominent role in Chapter 3 because the two-alternative forced-choice task ferrets were trained on required them to locate a foreground target amidst different spatial configurations of background noise (Fig. 3.1D). Spatial release from masking (SRM) refers to the improved intelligibility of a sound in noise when spatially separated (Stillman and Irwin, 1990; Peng et al., 2021). Though our results are incomplete and require further testing, by training the task with interleaved trials where the background masker was spatially paired or opposite of the target, we expected to see behavioral performance that suffered much more at more difficult SNRs in the monaural condition. Conversely, we expected in the binaural condition where target and distractor are spatially separate that SRM would improve target discriminability even at lower SNRs. Gathering a robust dataset

while recording from animals during these different spatial configurations and at more challenging SNRs would allow us to determine if spatial release from masking can be observed at the single-unit level.

### **4.3 Contributions to foreground response reduction outside of AC?**

My data and experiments focused almost entirely on the primary AC, briefly extending into secondary auditory cortex in Chapter 2, but how might the auditory system before it's stop in cortex affect my finding of foreground response reduction?

In Chapter 2, we considered the possibility of bottom-up processes from the peripheral auditory system may emphasize the spectrotemporal energy of backgrounds. To test this, we calculated spectral SNR to determine, within a neuron's receptive field, whether background or foreground had more energy. Although we did find a relationship between our relative gain metric of foreground response reduction and spectral SNR, the effect was not large enough to fully explain our result such that even when spectral SNR was 0 dB we still saw foreground response reduction (Fig. 2.8C). This led me to consider other response properties in lower-level areas that may contribute to the response interactions observed in A1.

For instance, the auditory periphery and midbrain contain numerous nonlinear response properties—saturation, phase locking, sensitivity to amplitude modulation—that can modify the signal reaching AC. Background noise poses a challenge because noise can cause saturation of firing or obscure phase locking. Auditory nerve fibers tuned to frequencies near peaks of the sound will produce a saturated response or one that phase locks to a single feature, producing weak fluctuations at the sound's fundamental. Meanwhile, fibers tuned away from the frequency action can be dominated by multiple harmonics to create slow fluctuations at the fundamental that get enhanced by inferior colliculus neurons tuned to amplitude modulation as bandpass, band-reject, low-pass, or high-pass (Carney et al., 2015; Kim et al., 2020). Some of these tunings are characterized by excitation, inhibition, or a mix to create a neural code in the midbrain that consequently improves the perception of more harmonic sounds like speech, or in our case, foregrounds (Fig. 2.6A). It is possible then that somewhere in this code lies a contributing factor to the relative reduction of foreground responses in A1 when paired with background noise. With a midbrain code sensitive to neural fluctuations resulting from sounds with harmonic structure, contrasts between these sounds and more “noisy” sounds can be enhanced, serving as an “edge-detector” (Carney et al., 2015) which could be useful in parsing the boundary between foreground and background.

Though this notion may seem like it hints at the beginning of foreground selectivity in sub-cortical areas which would be inconsistent with our findings of reduced foreground responses in A1, I instead argue it could fit with our hypothesis of noise invariance arising through the subtraction of population-

wide representations of noise. Having a code that can parse features of sound mixtures could allow for the enhanced representation of noise in the early AC that is necessary for its subtraction as the foreground representation emerges through secondary and tertiary AC.

Another property of lower-level like areas discussed in the Introduction (Section 1.2.1) was the spike timing codes brought about by the precise, low-latency firing of neurons in the inferior colliculus and, to a lesser extent, AC which encode information through precise timings of responses to sound features like onsets (Trussell, 1999). Also discussed was the relative predominance of rate codes in AC that represent information following amplitude modulations (Lu et al., 2001). In our study we chose to focus on rate codes—asking to what extent firing in response to a foreground is reduced when in the presence of a background. We chose to do so because rate codes have established to be relatively dominant in AC (Niwa et al., 2012; Bagur et al., 2025). Still, this does not necessarily preclude the possibility of spike timing codes playing a role in the AC representation of foregrounds in the presence of backgrounds. For instance, although we observed a dominance of firing relating to background sounds, information about the foreground could remain contained in a temporal code such that key onsets or features of the foreground remain encoded, albeit to a smaller magnitude of response. In this way, while our results show that responses to overlapping background and foreground sounds are dominated by a response to the background, if onset responses to temporally transient foreground sounds remain precise, this may be sufficient information about the foreground in a timing code. A future direction of the data presented in this dissertation could explore this possibility by analyzing responses to compare how and to what extent the timing of firing to foregrounds is changed in the presence of a background, potentially revealing a complementary role of spike timing code in the encoding of natural background/foreground contrasts.

#### **4.4 Behavioral consequences of foreground response reduction**

My results in our trained and behaving animals (Chapter 3) show the persistence of the foreground-specific response reduction observed in untrained, passively listening animals (Chapter 2), a significant control. Inherent to the stimulus paradigm of the passive experiments in Chapter 2 was an assumption that a background/foreground contrast produces streaming (Moore et al., 2013; Rabinowitz et al., 2013; Mesgarani et al., 2014) with an automatic perceptual preference towards narrowband, temporally dynamic foreground sounds even in untrained ferrets. It is also well described that top-down processing can enhance behaviorally relevant stimuli in humans (Ding and Simon, 2012; Mesgarani and Chang, 2012; O’Sullivan et al., 2013, 2015; Mesgarani et al., 2014). Taken together, a stronger representation of foreground responses in trained animals attending to foregrounds during a complex two-alternative



forced-choice task would make sense given top-down enhancement. It would also indicate a flaw in our assumptions about the salience of a foreground in the passive paradigm. However, we found comparable results between our passive and behaving recordings, allowing me to speculate more confidently as to the role of foreground response reduction in behavioral performance and auditory streaming perception.

### **Mechanisms shaping the neural representation of concurrent sounds**

Though our animals have achieved proficiency on an auditory streaming task (Fig. 3.1E), we have trials—especially at the more difficult SNRs—where the animal erred, either selecting the incorrect lickspout or providing no response at all. With a more robust dataset with a larger number of incorrect or null trials we could repeat our weighted gain analysis on these trials with behavioral performance as a readout of unsuccessful or incorrect streaming. In Chapter 2, I hypothesized that foreground response reduction in early stages of auditory processing may reflect a need to construct a distributed representation of distracting background stimuli to subtract this information from population-wide activity (McDermott and Simoncelli, 2011; Shamma et al., 2011; Tye et al., 2024). If this were true, the enhanced encoding of distracting stimuli in the early stages of auditory processing would be helpful for downstream responses in the behavioral output. As a result, unsuccessful streaming might then arise from an incomplete representation of distractors that would prevent the requisite downstream foreground acuity. In incorrect behavioral trials we may then expect to see less robust foreground response reduction, or less background dominance in the population, to reveal a functional role of reduced foreground responses. Where, then, would a foreground signal with noise representations subtracted become useful?

### **The emergence of foreground specificity**

A hypothesis where a noise robust signal emerges through the auditory hierarchy by subtracting population-wide representations of noise would require some level of modification through processing stages in AC. In my results showing foreground response reduction in secondary AC (Fig. 2.2C), foreground specificity clearly doesn't arise at this level. However, given the smaller magnitude of foreground reduction driven by a more rapid adaptation to background sounds in PEG (Fig. 2.4, 2.5C), the gradual emergence of foreground specificity seems tenable. Similarly, an fMRI study in humans found evidence that noise-invariance may not yet emerge at the level of A1 (Kell and McDermott, 2017), while at higher levels in the auditory hierarchy such as the superior temporal gyrus (STG), local field potential and fMRI have reported the kind of foreground specificity consistent with noise-robust perception (Kell

and McDermott, 2019; Khalighinejad et al., 2019). We now need to look forward in the ferret AC to hypothesize where invariance emerges as it does in human STG, a more high-level auditory field.

A study in the ferret Ventral Posterior field (VPr) gives insight into this tertiary area that is notably understudied due to its extremely lateral location (Elgueda et al., 2019). VPr was found to show selective enhancement of extracellular responses to behaviorally relevant target stimuli, an enhancement that is amplified during active performance of an auditory discrimination task. Importantly, this contrast between enhancement during active performance and passive listening gradually increased from recordings in A1 to PEG to VPr, finding enhancement to be weakly represented in secondary areas and even less so in A1. This result is reminiscent of results from Chapter 3 in A1 showing a minimal effect between foreground relative gain between active and passive trial blocks in our two-alternative forced-choice task (Fig. 3.2D). Were the emergence of target enhancement from A1 to VPr to have parallels to representations of our natural background/foreground sound contrasts, we might expect to see incrementally less response reduction of our target foregrounds in PEG and VPr, with increasing contrast between our behaving and passive recordings. Probing whether these predicted differences and our previously observed differences between A1 and PEG remain is a very accessible extension of our behavioral data given our current recording setup. Should we see these changes in PEG, it would certainly be worth attempting to extend our recording capability to higher auditory areas.

### **Decoding behavioral performance from neural responses**

Having a robust number of correct, incorrect, and null trials could also allow for decoding analyses to explore whether neural responses can predict performance. In the Introduction, I discussed encoding models that can predict neural responses based on the stimulus input after extensive learning of how that neuron responds to a large sampling of stimuli. Decoding models, on the other hand, allow us to look at neural responses and see what information is encoded in the response. Can we make a model, then, with an extensive neural dataset during correct and incorrect behavior that can predict behavioral performance based on neural response? More specifically, if we provide this model with information about the spatial orientation and SNR of the target and distractor and the accompanying neural responses on a particular trial, can it predict whether the animal chose correctly based on neural response? If so, this would indicate a neural signature of correct streaming or that unsuccessful streaming took place—the target was too quiet and thus no or uninformative information was encoded about it, the animal wasn't paying attention when a more challenging monaural task occurred and they didn't spatially resolve the location of the target having only one source, etc. Or, as discussed above, perhaps incorrect trials more frequently feature a smaller amount of foreground response reduction.

Indeed, neural decoding is a useful tool to understand the relationship between behavioral and neural data (Paninski et al., 2007; Glaser et al., 2020; Zhang et al., 2024), and data out of the lab gives us precedent for this kind of inference of perception using neural responses. Individuals with normal hearing tend to fuse binaural sounds when there is a relatively small (0.1-0.2 octave) pitch difference between the ears, while more disparately pitched sounds are perceived as two separate sounds. Perceptual evidence in humans show that when two synthetic vowel sounds are presented dichotically and separated by little to no frequency difference spectral averaging is likely to occur resulting in the percept of a distinct, intermediate vowel sound (Reiss and Molis, 2021). In our lab, a project seeks to determine the neural basis for this fusion by recording neurons of ferret AC as they passively listen to vowel presentations. Preliminary neural data shows that when two vowels are presented dichotically and with no frequency separation, neural responses tend to resemble responses to neither vowel alone, but rather the response resembles that of a third vowel sound that in the human study was indicative of binaural fusion. This suggests an averaging of neural responses and potentially the perception of this fused sound. It also means that in our streaming task we may be able to use neural responses and behavioral performance to infer whether streaming took place or not.

### **Non-auditory components of the auditory response**

We can also use our behavioral results to explore the presence of neurons in A1 that do not explicitly care about auditory information but seem to have a motor component to them (Fig. 3.3). On first thought, it seems unintuitive: neurons in the auditory system that don't care about auditory information—what are neurons that don't care about sounds doing in the sound area of the brain? On the contrary, hearing doesn't exist on its own in our experience of the world and ought to be informed by contexts that may reflect behavioral state or other sensory modalities. As such, recent studies have shown that neurons in sensory cortices indeed do encode information about place, body orientation, movement, etc. (Mertens et al., 2023; Mimica et al., 2023) which can serve as a multimodal reference or even to reflect expectation as in the suppression of self-generated noises (Holey and Schneider, 2023). These variables, namely position relative to sound source and movement can be accounted for to increase accuracy in encoding models.

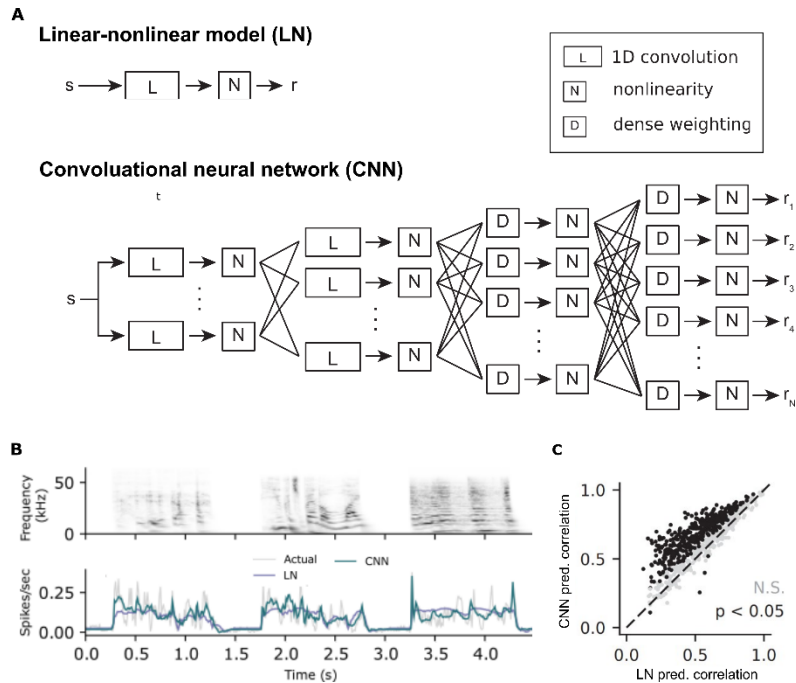
During free-moving behavior, we record animals with an overhead camera to track their location. As a result, we have data on the position of the animal during any given part of a stimulus presentation or behavioral outcome. We could, for instance, then plot the animal's two-dimensional position relative to the center lick port versus the PSTH of a neuron to see how fluctuations in head position relative to the speakers affect encoding of the sound stimuli. Or, maybe once the animal starts moving its mind is made up and it stops listening in favor of collecting the expected reward. As it relates to our non-auditory AC

neurons, knowing the movement of the animal at any given time can inform us what these neurons in the AC are *actually* encoding and allow us to speculate how that may be useful in the streaming task.

## 4.5 Modeling complex representations of natural sounds with deep learning

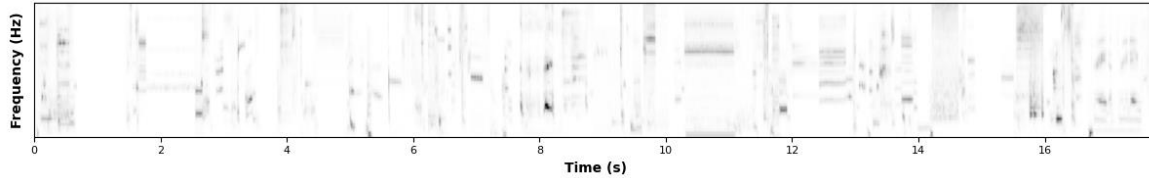
We have discussed (Section 1.2.5) the spectrotemporal receptive field (STRF) as a linear model that summarizes the unique combination of spectral, temporal, and level properties to which a given neuron is tuned (Aertsen and Johannesma, 1981). Once fit using a large number of responses from that neuron across a large sampling of statistically diverse stimuli, an STRF can do a remarkable job at predicting neural responses to sound stimuli as a linear feature detector (Thorson et al., 2015). A limitation of this model, though, is that progressively higher areas of AC are tuned to increasing complexity that may be hard to capture with this static model (Atencio et al., 2009). Particularly given the importance of context and changing spectrotemporal relationships within our paradigm of overlapping natural backgrounds and foregrounds, we also discussed the dynamic STRF (DSTRF). The DSTRF uses deep learning to generate an STRF that changes over time as the sound progresses (Keshishian et al., 2020), which could inform how evolving contexts affect a neuron’s tuning. Taking a step back, can we create a model capable of predicting the response interactions in complex natural sound mixtures? It seems like we can.

Convolutional neural networks (CNNs) are a deep learning architecture that can be used to model sensory processing using neurophysiology data (Butts, 2019; Richards et al., 2019). Though already well established in the world of the visual system to model natural image representation (McIntosh et al., 2017; Cadena et al., 2019), until recently it was not clear that CNNs could be used to characterize single-unit activity in AC due to these neurons having complex tuning that requires very large datasets. A recent study (Pennington and David, 2023) trained both an STRF-like linear model and a CNN population encoding model to predict the activity of large numbers of neurons during natural sound presentations that was generalizable to untrained units and never-before-seen sounds (Fig. 4.1A). The CNN model outperformed more traditional linear models (Fig. 4.1B).



**Figure 4.1.** Introduction to deep learning architectures as models of auditory encoding of natural sounds **A**, Schematics of model architectures showing a relatively less complex linear-nonlinear model (like the STRFs discussed in Fig. 1.17) and the more complex convolutional neural network. **B**, Example PSTH responses of a neuron in response to the above natural sounds as well as PSTHs predicted by the LN and CNN. **C**, Across most recordings, the CNN does a better job of predicting neural responses to complex natural sounds compared to the LN. Figures reproduced or adapted from Pennington & David, 2023.

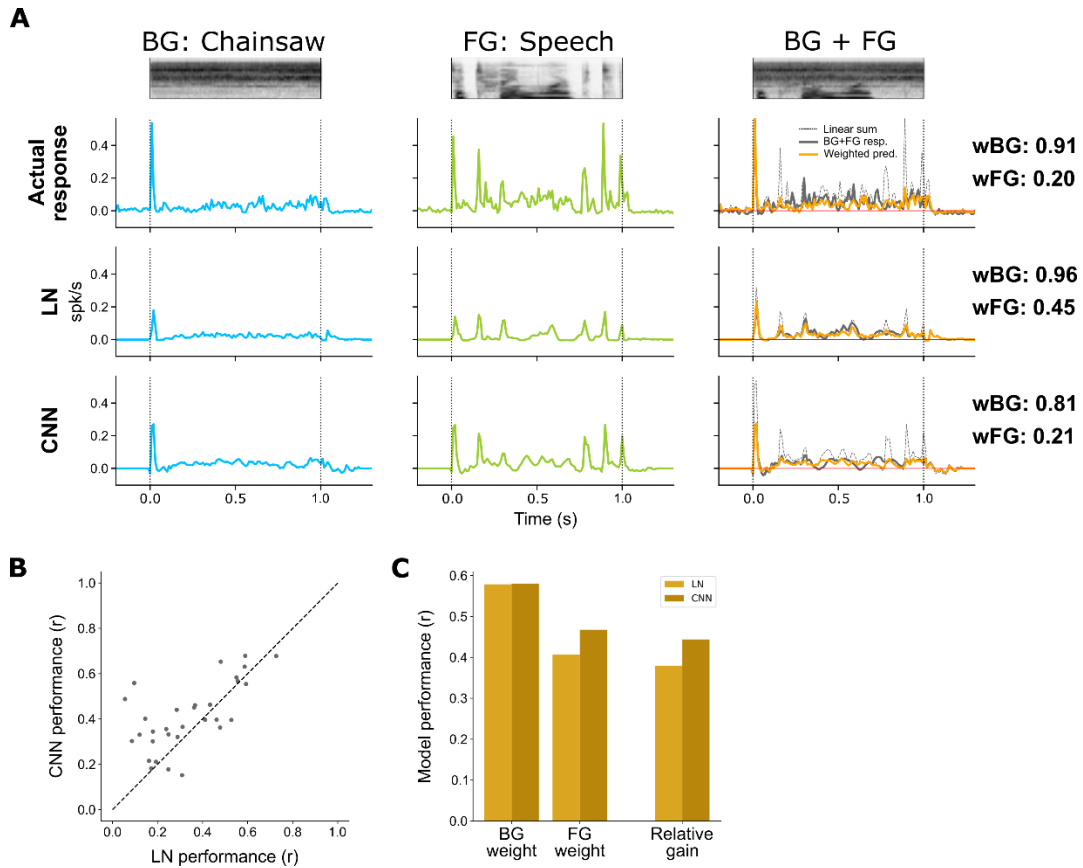
Though in preliminary stages, an extension of the work presented in this dissertation aims to model and predict the complex interactions we have observed and discussed that lead to the relative reduction of foreground responses detailed in Chapters 2 and 3. During several days of recordings in naïve, untrained animals that led to the data presented in Chapter 2, we also included trial blocks which presented thousands of 0.125-0.8 s natural sound excerpts (Fig. 4.2). Here the sound category (background or foreground) was not relevant and therefore not curated, though the set of sounds did include exemplars belonging to both categories. These large sets of natural sounds provided a huge and rapidly presented natural sound dataset on which we could train the CNN to predict responses to novel natural sounds as previously shown (Pennington and David, 2023). Consistent with this previous work, we found the CNN trained on these diverse natural sound sets able to accurately predict responses to novel, isolated 1 s backgrounds and foregrounds exemplars used throughout Chapters 2 and 3 (Fig. 4.3A, *left, middle*).



**Figure 4.2.** Example stimulus spectrogram of large, diverse natural sound set whose responses were used to train our models. Each natural sound excerpt, independent of background/foreground categorization were presented at short (0.125-0.8 s) durations.

We next sought to apply these models to our natural background/foreground stimuli contrasts to determine the extent to which they can accurately predict responses to concurrent presentations of natural sounds despite having never been trained on overlapping sounds with masked spectrotemporal properties. Indeed, both the LN and CNN were able to predict responses to BG+FG sound combinations (Fig. 4.3A, *right, black*). Further, fitting our weighted linear model (Chapter 2) to the LN and CNN-predicted individual BG and FG responses yielded similar weights and response to the BG+FG response (orange). The linear sum (dashed line) is also shown to describe how the CNN more accurately depicts the FG-specific response reduction observed in the neural response. Overall, the CNN outperformed the LN model in predicting BG and FG weights across most recording sites (Fig. 4.3B). The improved prediction accuracy of the CNN was shown to be from more accurately predicting lower FG weights leading to a more accurate RG than the LN, despite both being equally accurate with BG weights (Fig. 4.3C).

Together, these results reveal the versatility of deep learning as applied to the auditory system and, novelly, its ability to be applied to complex sound mixtures. Though this extension of my primary dissertation work is in its early stages, it is exciting to see my result of foreground response reduction that we—for years—have considered to be unexpected able to be recapitulated and captured by modelling. Extending tools such as the DSTRF (Keshishian et al., 2020) to increase the interpretability of CNNs by visualizing the activation patterns that drive the network output could allow us better understand how our model is arriving at this result of foreground response reduction. The linear function equivalent to the one the network applies to each repetition of a stimulus is called the locally linear receptive field and can be found at any moment in time and allow us to visualize how different sound contexts affect a neuron’s response and tuning. This visualization could give insight into the model-predicted foreground response reduction, representing an exciting next step in my dissertation research.



**Figure 4.3.** Models trained on individual natural sounds were able to predict neural responses to BG+FG sound combinations **A**, Example PSTHs comparing (rows) actual neural responses, LN prediction, and CNN prediction for BG, FG, and BG+FG natural sounds (columns). In the rightmost column, the BG+FG response (black), weighted prediction of the individual BG and FG responses (orange), and the linear sum (dashed) of the BG (blue) and FG (green) response are compared. The respective BG and FG weights for each row are shown at right. **B**, Comparison of LN and CNN performance predicting BG and FG weights at each recording site ( $n=42$ ). **C**, Summary of model performance for all unit and sound pair instances. The models predict comparable BG weights, but the CNN more correctly predicts lower FG weights, thus the CNN predicts the relative reduction of FG responses more accurately.

## References

- Aertsen AM, Johannesma PI (1981) The spectro-temporal receptive field. A functional characteristic of auditory neurons. *Biol Cybern* 42:133–143.
- Akeroyd MA, Carlyon RP, Deeks JM (2005) Can dichotic pitches form two streams? *J Acoust Soc Am* 118:977–981.
- Andreou L-V, Kashino M, Chait M (2011) The role of temporal regularity in auditory segregation. *Hear Res* 280:228–235.
- Anon (2022) How Do We Hear? | NIDCD. Available at: <https://www.nidcd.nih.gov/health/how-do-we-hear> [Accessed August 22, 2024].
- Atencio CA, Sharpee TO, Schreiner CE (2009) Hierarchical computation in the canonical auditory cortical circuit. *Proc Natl Acad Sci U S A* 106:21894–21899.
- Atiani S, David SV, Elgueda D, Locastro M, Radtke-Schuller S, Shamma SA, Fritz JB (2014) Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. *Neuron* 82:486–499.
- Attias H, Schreiner CE (1996) Temporal low-order statistics of natural sounds. In: *Proceedings of the 9th International Conference on Neural Information Processing Systems*, pp 27–33 NIPS’96. Cambridge, MA, USA: MIT Press.
- Attias H, Schreiner CE (1997) Temporal low-order statistics of natural sounds. In: *Advances in Neural Information Processing Systems*, pp 103–109. MIT Press.
- Bagur S, Bourg J, Kempf A, Tarpin T, Bergaoui K, Guo Y, Ceballo S, Schwenkgrub J, Verdier A, Puel JL, Bourien J, Bathellier B (2025) A spatial code for temporal information is necessary for efficient sensory learning. *Sci Adv* 11:eadr6214.
- Bajo VM, Nodal FR, Moore DR, King AJ (2010) The descending corticocollicular pathway mediates learning-induced auditory plasticity. *Nat Neurosci* 13:253–260.
- Bandyopadhyay S, Shamma SA, Kanold PO (2010) Dichotomy of functional organization in the mouse auditory cortex. *Nat Neurosci* 13:361–368.
- Bartlett EL (2013) The organization and physiology of the auditory thalamus and its role in processing acoustic features important for speech perception. *Brain Lang* 126:29–48.
- Bar-Yosef O, Nelken I (2007) The effects of background noise on the neural responses to natural sounds in cat primary auditory cortex. *Front Comput Neurosci* 1 Available at: <https://www.frontiersin.org/articles/10.3389/neuro.10.003.2007> [Accessed November 7, 2023].
- Bee MA, Klump GM (2004) Primitive Auditory Stream Segregation: A Neurophysiological Study in the Songbird Forebrain. *J Neurophysiol* 92:1088–1104.
- Bendor D, Wang X (2005) The neuronal representation of pitch in primate auditory cortex. *Nature* 436:1161–1165.



- Berg RE (2024) sound. Available at: <https://www.britannica.com/science/sound-physics> [Accessed September 2, 2024].
- Bernstein LR, Trahiotis C (1985) Lateralization of low-frequency, complex waveforms: the use of envelope-based temporal disparities. *J Acoust Soc Am* 77:1868–1880.
- Best V, Thompson ER, Mason CR, Kidd G (2013) Spatial release from masking as a function of the spectral overlap of competing talkers. *J Acoust Soc Am* 133:3677–3680.
- Biebel UW, Langner G (2002) Evidence for interactions across frequency channels in the inferior colliculus of awake chinchilla. *Hear Res* 169:151–168.
- Bizley J, King A (2009) Visual influences on ferret auditory cortex. *Hear Res* 258:55–63.
- Bizley JK, Cohen YE (2013) The what, where and how of auditory-object perception. *Nat Rev Neurosci* 14:693–707.
- Bizley JK, Nodal FR, Nelken I, King AJ (2005) Functional organization of ferret auditory cortex. *Cereb Cortex N Y N 1991* 15:1637–1653.
- Blake DT, Merzenich MM (2002) Changes of AI receptive fields with sound density. *J Neurophysiol* 88:3409–3420.
- Bondy J, Becker S, Bruce I, Trainor L, Haykin S (2004) A novel signal-processing strategy for hearing-aid design: neurocompensation. *Signal Process* 84:1239–1253.
- Bregman AS (1990) *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA, US: The MIT Press.
- Bregman AS (1994) *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.
- Bregman AS, Ahad PA, Crum PAC, O'Reilly J (2000) Effects of time intervals and tone durations on auditory stream segregation. *Percept Psychophys* 62:626–636.
- Butler BE, Lomber SG (2013) Functional and structural changes throughout the auditory system following congenital and early-onset deafness: implications for hearing restoration. *Front Syst Neurosci* 7:92.
- Butts DA (2019) Data-Driven Approaches to Understanding Visual Neuron Activity. *Annu Rev Vis Sci* 5:451–477.
- Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolia AS, Bethge M, Ecker AS (2019) Deep convolutional models improve predictions of macaque V1 responses to natural images. *Einhäuser W, ed. PLOS Comput Biol* 15:e1006897.
- Carlyon RP (2004) How the brain separates sounds. *Trends Cogn Sci* 8:465–471.
- Carney LH, Li T, McDonough JM (2015) Speech Coding in the Brain: Representation of Vowel Formants by Midbrain Neurons Tuned to Sound Fluctuations. *eNeuro* 2:ENEURO.0004-15.2015.
- Cherry EC (1953) Some Experiments on the Recognition of Speech, with One and with Two Ears. *J Acoust Soc Am* 25:975–979.

- Christison-Lagay KL, Cohen YE (2014) Behavioral correlates of auditory streaming in rhesus macaques. *Hear Res* 309:17–25.
- Cusack R, Roberts B (2000) Effects of differences in timbre on sequential grouping. *Percept Psychophys* 62:1112–1120.
- Darwin CJ (1997) Auditory grouping. *Trends Cogn Sci* 1:327–333.
- David SV, Fritz JB, Shamma SA (2012) Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc Natl Acad Sci* 109:2144–2149.
- Davis ZW, Dotson NM, Franken TP, Muller L, Reynolds JH (2023) Spike-phase coupling patterns reveal laminar identity in primate cortex Izquierdo A, Huguenard JR, Thiele A, eds. *eLife* 12:e84512.
- de la Mothe LA, Blumell S, Kajikawa Y, Hackett TA (2006) Cortical connections of the auditory cortex in marmoset monkeys: core and medial belt regions. *J Comp Neurol* 496:27–71.
- Dent ML, Martin AK, Flaherty MM, Neilans EG (2016) Cues for auditory stream segregation of birdsong in budgerigars and zebra finches: Effects of location, timing, amplitude, and frequency. *J Acoust Soc Am* 139:674–683.
- Deshpande S, C. Sajjan S, Pujar H (2019) System to Transform Sound Energy Into Electricity. Available at: <https://papers.ssrn.com/abstract=3492946> [Accessed July 22, 2024].
- Deutsch D (1974) An auditory illusion. *Nature* 251:307–309.
- Diamond DM, Weinberger NM (1986) Classical conditioning rapidly induces specific changes in frequency receptive fields of single neurons in secondary and ventral ectosylvian auditory cortical fields. *Brain Res* 372:357–360.
- Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci* 109:11854–11859.
- Du J, Blanche TJ, Harrison RR, Lester HA, Masmanidis SC (2011) Multiplexed, High Density Electrophysiology with Nanofabricated Neural Probes. *PLOS ONE* 6:e26204.
- Efron B, Tibshirani R (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat Sci* 1:54–75.
- Elgueda D, Duque D, Radtke-Schuller S, Yin P, David SV, Shamma SA, Fritz JB (2019) State-dependent encoding of sound and behavioral meaning in a tertiary region of the ferret auditory cortex. *Nat Neurosci* 22:447–459.
- Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma SA (2009) Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron* 61:317–329.
- Eliades SJ, Tsunada J (2019) Chapter 25 - Marmosets in Auditory Research. In: *The Common Marmoset in Captivity and Biomedical Research* (Marini R, Wachtman L, Tardif S, Mansfield K, Fox J, eds), pp 451–475 American College of Laboratory Animal Medicine. Academic Press. Available at: <http://www.sciencedirect.com/science/article/pii/B978012811829000025X> [Accessed February 26, 2020].

- Escabí MA, Read HL (2003) Representation of spectrotemporal sound information in the ascending auditory pathway. *Biol Cybern* 89:350–362.
- Fay RR (1998) Auditory stream segregation in goldfish (*Carassius auratus*). *Hear Res* 120:69–76.
- Feng L, Wang X (2017) Harmonic template neurons in primate auditory cortex underlying complex sound processing. *Proc Natl Acad Sci U S A* 114:E840–E848.
- Fettiplace R (2020) Diverse Mechanisms of Sound Frequency Discrimination in the Vertebrate Cochlea. *Trends Neurosci* 43:88–102.
- Fishman YI, Reser DH, Arezzo JC, Steinschneider M (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hear Res* 151:167–187.
- Franken TP, Bondy BJ, Haimes DB, Goldwyn JH, Golding NL, Smith PH, Joris PX (2021) Glycinergic axonal inhibition subserves acute spatial sensitivity to sudden increases in sound intensity. *eLife* 10:e62183.
- Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci* 6:1216–1223.
- Furukawa S, Xu L, Middlebrooks JC (2000) Coding of sound-source location by ensembles of cortical neurons. *J Neurosci Off J Soc Neurosci* 20:1216–1228.
- Glaser JI, Benjamin AS, Chowdhury RH, Perich MG, Miller LE, Kording KP (2020) Machine Learning for Neural Decoding. *eNeuro* 7:ENEURO.0506-19.2020.
- Goense JBM, Feng AS (2012) Effects of noise bandwidth and amplitude modulation on masking in frog auditory midbrain neurons. *PloS One* 7:e31589.
- Griffiths TD, Warren JD (2004) What is an auditory object? *Nat Rev Neurosci* 5:887–892.
- Gruters KG, Groh JM (2012) Sounds and beyond: multisensory and other non-auditory signals in the inferior colliculus. *Front Neural Circuits* 6:96.
- Hackett TA (2011) Information flow in the auditory cortical network. *Hear Res* 271:133–146.
- Hamersky GR, Shaheen LA, Espejo ML, Wingert JC, David SV (2023) Unexpected suppression of neural responses to natural foreground versus background sounds in auditory cortex. :2023.11.20.567922 Available at: <https://www.biorxiv.org/content/10.1101/2023.11.20.567922v1> [Accessed August 13, 2024].
- Heller CR, Hamersky GR, David SV (2023) Task-specific invariant representation in auditory cortex. *eLife* 12 Available at: <https://elifesciences.org/reviewed-preprints/89936> [Accessed October 30, 2023].
- Holey BE, Schneider DM (2023) Sensation and expectation are embedded in mouse motor cortical activity. *bioRxiv*:2023.09.13.557633.
- Hui GK, Wong KL, Chavez CM, Leon MI, Robin KM, Weinberger NM (2009) Conditioned tone control of brain reward behavior produces highly specific representational gain in the primary auditory cortex. *Neurobiol Learn Mem* 92:27–34.

- Hulse SH, MacDougall-Shackleton SA, Wisniewski AB (1997) Auditory scene analysis by songbirds: stream segregation of birdsong by European starlings (*Sturnus vulgaris*). *J Comp Psychol Wash DC* 111:3–13.
- Itatani N, Klump GM (2017) Animal models for auditory streaming. *Philos Trans R Soc Lond B Biol Sci* 372.
- Izumi A (2002) Auditory stream segregation in Japanese monkeys. *Cognition* 82:B113-122.
- Jun JJ et al. (2017) Fully Integrated Silicon Probes for High-Density Recording of Neural Activity. *Nature* 551:232–236.
- Kato HK, Asinof SK, Isaacson JS (2017) Network-Level Control of Frequency Tuning in Auditory Cortex. *Neuron* 95:412-423.e4.
- Katsiamis AG, Drakakis EM, Lyon RF (2007) Practical Gammatone-Like Filters for Auditory Processing. *EURASIP J Audio Speech Music Process* 2007:1–15.
- Kell AJ, McDermott J (2017) Robustness to real-world background noise increases between primary and non-primary human auditory cortex. *J Acoust Soc Am* 141:3896–3896.
- Kell AJE, McDermott JH (2019) Invariance to background noise as a signature of non-primary auditory cortex. *Nat Commun* 10:3958.
- Keshishian M, Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2020) Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *eLife* 9:e53445.
- Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2019) Adaptation of the human auditory cortex to changing background noise. *Nat Commun* 10:2509.
- Kikuchi Y, Horwitz B, Mishkin M, Rauschecker JP (2014) Processing of harmonics in the lateral belt of macaque auditory cortex. *Front Neurosci* 8 Available at: <https://www.frontiersin.org/articles/10.3389/fnins.2014.00204> [Accessed November 5, 2023].
- Kim DO, Carney L, Kuwada S (2020) Amplitude modulation transfer functions reveal opposing populations within both the inferior colliculus and medial geniculate body. *J Neurophysiol* 124:1198–1215.
- King A, Middlebrooks J (2010) Cortical Representation of Auditory Space. In: *The Auditory Cortex*, pp 329–341.
- Kisley MA, Gerstein GL (2001) Daily variation and appetitive conditioning-induced plasticity of auditory cortex receptive fields. *Eur J Neurosci* 13:1993–2003.
- Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. *J Comput Neurosci* 9:85–111.
- Kline AM, Aponte DA, Kato HK (2023) Distinct nonlinear spectrotemporal integration in primary and secondary auditory cortices. *Sci Rep* 13:7658.

- Kline AM, Aponte DA, Tsukano H, Giovannucci A, Kato HK (2021) Inhibitory gating of coincidence-dependent sensory binding in secondary auditory cortex. *Nat Commun* 12:4610.
- Lee CC (2013) Thalamic and cortical pathways supporting auditory processing. *Brain Lang* 126:22–28.
- Lesica NA, Grothe B (2008) Efficient temporal processing of naturalistic sounds. *PloS One* 3:e1655.
- López Espejo M, David SV (2024) A sparse code for natural sound context in auditory cortex. *Curr Res Neurobiol* 6:100118.
- Lu T, Liang L, Wang X (2001) Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat Neurosci* 4:1131–1138.
- Lyon RF, Katsiamis AG, Drakakis EM (2010) History and future of auditory filter models. In: *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp 3809–3812 Available at: <https://ieeexplore.ieee.org/document/5537724> [Accessed July 30, 2024].
- Ma L, Yin P, Micheyl C, Oxenham AJ, Shamma SA (2010) Behavioral Measures of Auditory Streaming in Ferrets (*Mustela putorius*). *J Comp Psychol Wash DC* 1983 124:317–330.
- Maier A, Adams G, Aura C, Leopold D (2010) Distinct Superficial and Deep Laminar Domains of Activity in the Visual Cortex during Rest and Stimulation. *Front Syst Neurosci* 4 Available at: <https://www.frontiersin.org/articles/10.3389/fnsys.2010.00031> [Accessed October 30, 2023].
- Malone BJ, Heiser MA, Beitel RE, Schreiner CE (2017) Background noise exerts diverse effects on the cortical encoding of foreground sounds. *J Neurophysiol* 118:1034–1054.
- Maor I, Shwartz-Ziv R, Feigin L, Elyada Y, Sompolinsky H, Mizrahi A (2019) Neural Correlates of Learning Pure Tones or Natural Sounds in the Auditory Cortex. *Front Neural Circuits* 13:82.
- McDermott JH (2009) The cocktail party problem. *Curr Biol CB* 19:R1024-1027.
- McDermott JH, Oxenham AJ (2008) Spectral completion of partially masked sounds. *Proc Natl Acad Sci U S A* 105:5939–5944.
- McDermott JH, Simoncelli EP (2011) Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron* 71:926–940.
- McIntosh LT, Maheswaranathan N, Nayebi A, Ganguli S, Baccus SA (2017) Deep Learning Models of the Retinal Response to Natural Scenes.
- McPherson MJ, Grace RC, McDermott JH (2022) Harmonicity aids hearing in noise. *Atten Percept Psychophys* 84:1016–1042.
- MED-EL (2017) Why Long Electrode Arrays: More Natural Tonotopic Coding. *MED-EL Prof Blog* Available at: <https://blog.medel.pro/products-updates/natural-tonotopic-coding/> [Accessed July 22, 2024].
- Mehta AH, Jacoby N, Yasin I, Oxenham AJ, Shamma SA (2017) An auditory illusion reveals the role of streaming in the temporal misallocation of perceptual objects. *Philos Trans R Soc Lond B Biol Sci* 372:20160114.

- Mendoza-Halliday D, Major AJ, Lee N, Lichtenfeld M, Carlson B, Mitchell B, Meng PD, Xiong Y (Sophy), Westerberg JA, Maier A, Desimone R, Miller EK, Bastos AM (2023) A ubiquitous spectrolaminar motif of local field potential power across the primate cortex. :2022.09.30.510398 Available at: <https://www.biorxiv.org/content/10.1101/2022.09.30.510398v4> [Accessed January 12, 2024].
- Mertens PEC, Marchesi P, Ruikes TR, Oude Lohuis M, Krijger Q, Pennartz CMA, Lansink CS (2023) Coherent mapping of position and head direction across auditory and visual cortex. *Cereb Cortex* 33:7369–7385.
- Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–236.
- Mesgarani N, David SV, Fritz JB, Shamma SA (2014) Mechanisms of noise robust representation of speech in primary auditory cortex. *Proc Natl Acad Sci U S A* 111:6792–6797.
- Micheyl C, Hunter C, Oxenham AJ (2010) Auditory stream segregation and the perception of across-frequency synchrony. *J Exp Psychol Hum Percept Perform* 36:1029–1039.
- Middlebrooks JC, Onsan ZA (2012) Stream segregation with high spatial acuity. *J Acoust Soc Am* 132:3896–3911.
- Middlebrooks JC, Pettigrew JD (1981) Functional classes of neurons in primary auditory cortex of the cat distinguished by sensitivity to sound location. *J Neurosci Off J Soc Neurosci* 1:107–120.
- Miller LM, Escabí MA, Read HL, Schreiner CE (2002) Spectrotemporal Receptive Fields in the Lemniscal Auditory Thalamus and Cortex. *J Neurophysiol* 87:516–527.
- Mimica B, Tombaz T, Battistin C, Fuglstad JG, Dunn BA, Whitlock JR (2023) Behavioral decomposition reveals rich encoding structure employed across neocortex in rats. *Nat Commun* 14:3947.
- Mischler G, Keshishian M, Bickel S, Mehta AD, Mesgarani N (2023) Deep neural networks effectively model neural adaptation to changing background noise and suggest nonlinear noise filtering methods in auditory cortex. *NeuroImage* 266:119819.
- Mishra AP, Harper NS, Schnupp JWH (2021) Exploring the distribution of statistical feature parameters for natural sound textures. *PLOS ONE* 16:e0238960.
- Młynarski W, McDermott JH (2019) Ecological origins of perceptual grouping principles in the auditory system. *Proc Natl Acad Sci* 116:25355–25364.
- Montes-Lourido P, Kar M, David SV, Sadagopan S (2021) Neuronal selectivity to complex vocalization features emerges in the superficial layers of primary auditory cortex. *PLOS Biol* 19:e3001299.
- Moore BCJ, Gockel HE (2012) Properties of auditory stream formation. *Philos Trans R Soc B Biol Sci* 367:919–931.
- Moore RC, Lee T, Theunissen FE (2013) Noise-invariant neurons in the avian auditory cortex: hearing the song in noise. *PLoS Comput Biol* 9:e1002942.

- Narayan R, Best V, Ozmeral E, McClaine E, Dent M, Shinn-Cunningham B, Sen K (2007) Cortical interference effects in the cocktail party problem. *Nat Neurosci* 10:1601–1607.
- Nelken I, Rotman Y, Yosef OB (1999) Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* 397:154–157.
- Ni R, Bender DA, Shanechi AM, Gamble JR, Barbour DL (2017) Contextual effects of noise on vocalization encoding in primary auditory cortex. *J Neurophysiol* 117:713–727.
- Niwa M, Johnson JS, O’Connor KN, Sutter ML (2012) Active engagement improves primary auditory cortical neurons’ ability to discriminate temporal modulation. *J Neurosci Off J Soc Neurosci* 32:9323–9334.
- Noda T, Takahashi H (2019) Behavioral evaluation of auditory stream segregation in rats. *Neurosci Res* 141:52–62.
- Norman-Haignere SV, McDermott JH (2018) Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biol* 16 Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6292651/> [Accessed April 2, 2020].
- Osmanski MS, Song X, Wang X (2013) The Role of Harmonic Resolvability in Pitch Perception in a Vocal Nonhuman Primate, the Common Marmoset (*Callithrix jacchus*). *J Neurosci* 33:9161–9168.
- Osmanski MS, Wang X (2011) Measurement of absolute auditory thresholds in the common marmoset (*Callithrix jacchus*). *Hear Res* 277:127–133.
- O’Sullivan JA, Crosse MJ, Power AJ, Lalor EC (2013) The effects of attention and visual input on the representation of natural speech in EEG. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf* 2013:2800–2803.
- O’Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cereb Cortex N Y N 1991* 25:1697–1706.
- Overath T, Kumar S, von Kriegstein K, Griffiths TD (2008) Encoding of Spectral Correlation over Time in Auditory Cortex. *J Neurosci* 28:13268–13273.
- Pachitariu M, Steinmetz N, Kadir S, Carandini M, Harris K (2016) Fast and accurate spike sorting of high-channel count probes with KiloSort.
- Paninski L, Pillow J, Lewi J (2007) Statistical models for neural encoding, decoding, and optimal stimulus design. *Prog Brain Res* 165:493–507.
- Peng ZE, Pausch F, Fels J (2021) Spatial release from masking in reverberation for school-age children. *J Acoust Soc Am* 150:3263.
- Pennington JR, David SV (2023) A convolutional neural network provides a generalizable model of natural sound coding by neural populations in auditory cortex. *PLOS Comput Biol* 19:e1011110.

- Popham S, Boebinger D, Ellis DPW, Kawahara H, McDermott JH (2018) Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nat Commun* 9:2122.
- Purves D, Augustine GJ, Fitzpatrick D, Katz LC, LaMantia A-S, McNamara JO, Williams SM (2001) The Auditory Cortex. In: *Neuroscience*. 2nd edition. Sinauer Associates. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK10900/> [Accessed March 19, 2024].
- Qiu A, Schreiner CE, Escabi MA (2003) Gabor Analysis of Auditory Midbrain Receptive Fields: Spectro-Temporal and Binaural Composition. *J Neurophysiol* 90:456–476.
- Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH (2013) Constructing noise-invariant representations of sound in the auditory pathway. *PLoS Biol* 11:e1001710.
- Rauschecker JP, Tian B, Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268:111–114.
- Reiss LAJ, Molis MR (2021) An Alternative Explanation for Difficulties with Speech in Background Talkers: Abnormal Fusion of Vowels Across Fundamental Frequency and Ears. *J Assoc Res Otolaryngol JARO* 22:443–461.
- Richards BA et al. (2019) A deep learning framework for neuroscience. *Nat Neurosci* 22:1761–1770.
- Saderi D, Buran BN, David SV (2020) Streaming of repeated noise in primary and secondary fields of auditory cortex. *bioRxiv:738583*.
- Schaefer MK, Hechavarría JC, Kössl M (2015) Quantification of mid and late evoked sinks in laminar current source density profiles of columns in the primary auditory cortex. *Front Neural Circuits* 9 Available at: <https://www.frontiersin.org/articles/10.3389/fncir.2015.00052> [Accessed October 30, 2023].
- Schwartz ZP, David SV (2018) Focal Suppression of Distractor Sounds by Selective Attention in Auditory Cortex. *Cereb Cortex N Y N* 1991 28:323–339.
- Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* 34:114–123.
- Shearer DE, Molis MR, Bennett KO, Leek MR (2018) Auditory stream segregation of iterated rippled noises by normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 143:378.
- Shinn-Cunningham BG (2005) Influences of spatial cues on grouping and understanding sound.
- Siegle JH, López AC, Patel YA, Abramov K, Ohayon S, Voigts J (2017) Open Ephys: an open-source, plugin-based platform for multichannel electrophysiology. *J Neural Eng* 14:045003.
- Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am* 114:3394–3411.
- Slee SJ, David SV (2015) Rapid Task-Related Plasticity of Spectrotemporal Receptive Fields in the Auditory Midbrain. *J Neurosci* 35:13090–13102.
- Sollini J, Poole KC, Blauth-Muszkowski D, Bizley JK (2022) The role of temporal coherence and temporal predictability in the build-up of auditory grouping. *Sci Rep* 12:14493.



- Song X, Osmanski MS, Guo Y, Wang X (2016) Complex pitch perception mechanisms are shared by humans and a New World monkey. *Proc Natl Acad Sci* 113:781–786.
- Stecker GC, Gallun F (2012) Binaural Hearing, Sound Localization, and Spatial Hearing.
- Stecker GC, Middlebrooks JC (2003) Distributed coding of sound locations in the auditory cortex. *Biol Cybern* 89:341–349.
- Stillman JA, Irwin R (1990) Signal detectability in the presence of monotic or dichotic noise bands of equal or unequal levels. *Percept Psychophys* 47 Available at: <https://pubmed.ncbi.nlm.nih.gov/2326150/> [Accessed September 3, 2024].
- Suga N (1995) Sharpening of frequency tuning by inhibition in the central auditory system: tribute to Yasuji Katsuki. *Neurosci Res* 21:287–299.
- Sussman ES, Horváth J, Winkler I, Orr M (2007) The role of attention in the formation of auditory streams. *Percept Psychophys* 69:136–152.
- Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Netw Bristol Engl* 12:289–316.
- Theunissen FE, Elie JE (2014) Neural processing of natural sounds. *Nat Rev Neurosci* 15:355–366.
- Theunissen FE, Sen K, Doupe AJ (2000) Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds. *J Neurosci* 20:2315–2331.
- Thorson IL, Liénard J, David SV (2015) The Essential Complexity of Auditory Receptive Fields. *PLOS Comput Biol* 11:e1004628.
- Tollin DJ (2003) The lateral superior olive: a functional role in sound source localization. *Neurosci Rev J Bringing Neurobiol Neurol Psychiatry* 9:127–143.
- Trainito C, von Nicolai C, Miller EK, Siegel M (2019) Extracellular Spike Waveform Dissociates Four Functionally Distinct Cell Classes in Primate Cortex. *Curr Biol* 29:2973-2982.e5.
- Trussell LO (1999) Synaptic mechanisms for coding timing in auditory neurons. *Annu Rev Physiol* 61:477–496.
- Tye KM, Miller EK, Taschbach FH, Benna MK, Rigotti M, Fusi S (2024) Mixed selectivity: Cellular computations for complexity. *Neuron* 112:2289–2303.
- Warren RM, Bashford JA, Lenz PW (2013) How Broadband Speech May Avoid Neural Firing Rate Saturation at High Intensities and Maintain Intelligibility. *Proc Meet Acoust Acoust Soc Am* 13:3426.
- Winkler I, Denham SL, Nelken I (2009) Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends Cogn Sci* 13:532–540.
- Zhai X, Khatami F, Sadeghi M, He F, Read HL, Stevenson IH, Escabí MA (2020) Distinct neural ensemble response statistics are associated with recognition and discrimination of natural sound textures. *Proc Natl Acad Sci* 117:31482–31493.

Zhang LI, Bao S, Merzenich MM (2001) Persistent and specific influences of early acoustic environments on primary auditory cortex. *Nat Neurosci* 4:1123–1130.

Zhang Y, Lyu H, Hurwitz C, Wang S, Findling C, Hubert F, Pouget A, Varol E, Paninski L (2024) Exploiting correlations across trials and behavioral sessions to improve neural decoding. [bioRxiv:2024.09.14.613047](https://doi.org/10.1101/2024.09.14.613047).