Behavioral and neural selectivity for acoustic signatures of vocalizations

Lam Tsz Nina So

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Behavioral and neural selectivity for acoustic signatures of vocalizations

Lam Tsz Nina So

Vocal communication relies on the ability of listeners to identify, process, and respond to vocal sounds produced by others in complex environments. In order to accurately recognize these signals, animals' auditory systems must robustly represent acoustic features that distinguish vocal sounds from other environmental sounds. In this dissertation, I describe experiments combining acoustic, behavioral, and neurophysiological approaches to identify behaviorally relevant vocalization features and understand how they are represented in the brain. First, I show that vocal responses to communication sounds in songbirds depend on the presence of specific spectral signatures of vocalizations. Second, I identify an anatomically localized neural population in the auditory cortex that shows selective responses for behaviorally relevant sounds. Third, I show that these neurons' spectral selectivity is robust to acoustic context, indicating that they could function as spectral signature detectors in a variety of listening conditions. Last, I deconstruct neural selectivity for behaviorally relevant sounds and show that it is driven by a sensitivity to deep fluctuations in power along the sound frequency spectrum. Together, these results show that the processing of behaviorally relevant spectral features engages a specialized neural population in the auditory cortex, and elucidate an acoustic driver of vocalization selectivity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis advisor, Sarah Woolley, for her mentorship, time, and patience for the past five years. This work would not have been possible without her guidance, encouragement of my scientific curiosities, and trust in giving me the opportunity to be a part of the Woolley lab.

I would also like to thank Darcy Kelley, Nima Mesgarani, and Nate Sawtell for their thoughtful contributions over the past five years as members of my thesis committee. Their input has grounded and guided me through my scientific endeavors. Thank you to Dan Sanes, who has kindly provided his input by acting as my external committee member. Also, thank you to Shihab Shamma for sharing his insights on the analysis and interpretation of this work, and to Wes Grueber for reading and helping to improve this work.

During my time at Columbia, I had the fortune to intersect with and learn from lab mates and friends. I am grateful to Jordan Moore and Lana Rosis, who have been to me a combination of teachers, friends, and colleagues. Thank you for the sound advice, countless conversations, and afternoon coffees we've shared. Five years ago, I joined the graduate program with a wonderful group of scientists – Daniel Iascone, Georgia Pierce, Abby Russo, Patrick Stinson, and Claire Warriner. Thank you for your friendship and contagious passion for what you do. Special thanks to Georgia for letting me partake in life with her lovely cat, Tina.

Thank you to Patrick Thompson – my best friend and the most supportive, kind, and understanding partner that I could ask for – for your emotional support, mathematical consultations, and numerous other contributions. Lastly, to my parents, Lawrence and Sue – thank you for believing in me, and for your unconditional support since the beginning.

Chapter 1

INTRODUCTION

1.1 ABSTRACT

Vocal communication relies on the ability of listeners to detect, process, and respond appropriately to others' vocal sounds. A thorough understanding of this process involves characterizing the acoustics of vocal sounds, understanding listeners' responses to these sounds, and identifying neural mechanisms that support vocalization processing. This chapter provides an overview of current literature on the generation, perception, and neural processing of vocalization-typical acoustic features, with a particular focus on acoustic structure in the frequency domain.

Vocal acoustic signatures seen in animals across taxa result from shared vocal production mechanisms involving periodically oscillating sound sources. The perception of spectral features characterizing vocalizations has been most extensively studied in humans. Psychophysical and speech perception studies have revealed that spectral structure is important for pitch perception, sound segregation, and extraction of social information from voices. Studies in other animals suggest that sensitivity to vocalization-typical spectral features may be a shared attribute among vocal communicators.

In the realm of auditory neuroscience, much progress has been made in identifying brain structures and neural populations that represent and process complex spectral features. These studies commonly utilize synthetic sounds, leaving open the question of whether the principles derived apply to the processing of natural vocalizations. Understanding how sensory processing supports social communication will require the combination of neural and behavioral approaches,

as well as the design and use of stimuli that capture and isolate behaviorally salient parameters of complex vocalizations.

1.2 ACOUSTIC HALLMARKS OF VOCALIZATIONS

1.2.1 Acoustic signals function in social communication

Vocalizations are used for communication by animals ranging from mammals (Ehret & Riecke, 2002; Eliades & Miller, 2016) to birds (Brainard & Doupe, 2013) and frogs (Kelley, 2004). These acoustic signals serve critical social functions, including mate choice (Holveck & Riebel, 2007), parental care (Ehret & Riecke, 2002), territory and resource defense (Hall, 2004), species recognition (Charrier & Sturdy, 2005), and individual recognition (Rendall et al., 1996; Vignal et al., 2008). For human beings, perceiving others' vocal sounds is also an important component of social interaction. The verbal content of speech allows us to recognize others' intentions, thoughts, and ideas. The vocal quality of speech, outside of verbal comprehension, can convey emotional states (Thompson & Balkwill, 2006).

1.2.2 Common vocal production mechanisms result in shared vocal acoustics across species

While remarkable diversity in vocal sounds is found across taxa, shared vocalization production mechanisms result in common acoustic signatures. Vocal sounds of tetrapods originate from vocal fold oscillations resulting in the production of harmonic sounds. Harmonic sounds contain energy at integer multiples of a fundamental frequency (F0), leading to the appearance of evenly spaced frequency components along the linear spectral axis. For example, a harmonic sound with F0 of 500 Hz can contain energy at 500 Hz, 1000 Hz, and 1500 Hz as the three lowest frequency components. A natural consequence of harmonicity is a non-uniform distribution of

energy across frequencies; distinct peaks and valleys can be observed in the spectra of harmonic sounds. Deep modulations in the spectral profile distinguishes harmonic sounds from flat spectra sounds such as white noise. Harmonic and spectrally modulated sounds are found in the vocal repertoires of many birds and mammals; examples include songbirds (Elie & Theunissen, 2016), chickens (Marler, 2004), pigeons (Alonso et al., 2016), rodents (Ehret & Riecke, 2002; Fernández-vargas & Johnston, 2015), cats (Shipley et al., 2005), elephants (Soltis, 2010), humpback whales (Cazau et al., 2016), and non-human primates (Eliades & Miller, 2016; Kikuchi et al., 2014). Below I provide an overview of vocal production in mammals and birds.

Most current knowledge on mammalian vocal production has stemmed from studies of the physics of human vocal production. The mammalian vocal organ is the larynx, which is positioned at the top of the trachea. The larynx contains a pair of vocal folds composed of mucous membranes. The harmonic structure and F0 of human vocal sounds rely on two types of vocal-fold movement. First, the alternation between abduction (the separation of vocal folds) and adduction (bringing vocal folds together at the midline) controls whether airflow causes the vocal folds to vibrate. Abduction allows the production of unvoiced (aperiodic) sounds which lack harmonic structure. Adduction brings vocal folds into the air stream, thus allowing them to vibrate in the presence of airflow and resulting in the production of voiced (periodic) sounds containing harmonic structure. Second, during adduction, the stretching and relaxing of vocal folds lead to higher or lower periodicities of vibrations. The periodicity of vibration controls the F0 of the resulting vocal sound. (Belyk & Brown, 2017).

The syrinx, the avian vocal organ located at the base of the trachea where it branches into the bronchi, is functionally analogous to the mammalian larynx and supports a similar vocal production mechanism. The syrinx contains two pairs of labia (a medial and a lateral labium),

which function similarly to the vocal folds in humans. During sound production, these labia are set into vibration by passing air. The F0 of labial vibrations determine the F0 of sound, and the range of F0s are limited by the material properties of the labia. Changes in labial configuration, analogous to abduction and adduction observed in human vocal production, are assumed to occur in models of bird vocal production, though they have not been empirically demonstrated. The activity of syringeal muscles are thought to control the rate of labial vibration and thus impact the F0 of vocal sounds generated (Elemans, 2014; Riede & Goller, 2010).

The harmonic sounds generated by both mammalian and avian sound sources are subject to subsequent filtering to emphasize or de-emphasize certain frequencies. During human speech production, the position of the tongue, lips, and jaw can serve to suppress certain frequency components composing the harmonic sound. In birds, filtering can occur by controlling the vocal tract length, beak gape, and regulating the volume of the oropharyngeal-esophageal cavity (Elemans, 2014).

1.2.3 Different modes of vocal fold coupling generate diverse vocal acoustics

While vocal sounds are typically harmonic, resulting from periodic vocal-fold vibrations, in some instances, humans and other animals also produce noisy sounds. Noisy sounds are characterized by broadband spectra with energy at many different frequencies, and are sometimes described as rough and harsh-sounding. This acoustic structure is observed in human infant cries and in the vocalizations of adults with voice disorders, as well as in the vocalization repertoires of nonhuman animals such as rhesus macaques, piglets, and domestic dogs (Fitch et al., 2002; Owren, 2002). Noisy vocalizations are thought to result from intrinsic properties of the vocal production system, which is capable of creating highly complex and variable acoustic output from relatively simple neural commands.

The vocal folds can couple their oscillations in various ways to result in diverse acoustic outputs. Three modes of coupling are reviewed here. First, in the "standard" state, the vocal folds synchronize their vibrations at the same frequency, leading to periodic oscillations that result in the production of harmonic sounds. Second, in the "subharmonic" state, the two vocal folds can be synchronized but have different vibratory frequencies. This could result in the appearance of additional frequency components termed "subharmonics" in the resulting sound signal. Third, in the "chaotic" state, desynchronized coupled oscillators can result in aperiodic vibrations. This results in noisy vocal sounds that have wide regions of broadband energy, though they generally do not have equal energy at all frequencies as does white noise (Fitch et al., 2002).

It has been shown that vocal folds can rapidly transition between different oscillatory states, such as from the "standard" state to the "chaotic" state. Zebra finches (*Taeniopygia guttata)*, for example, can produce continuous sound elements with harmonic segments that are directly followed by noisy segments without any period of silence in between. In an *in vitro* preparation of the syrinx, continuous variation in simple control parameters caused sudden jumps in acoustic output, such that vocal signals could transition rapidly from harmonic to noisy. Hence, fast intra-syllable transitions between harmonic and noisy sounds can be at least partially attributed to intrinsic dynamics of the syrinx (Fee et al., 1998).

1.3 PERCEPTUAL AND BEHAVIORAL IMPORTANCE OF VOCALIZATION-TYPICAL SPECTRAL FEATURES IN HUMANS

Vocalizations are a behaviorally important category of sounds, containing social information beneficial for animals' survival and reproduction. Hence, it is important to understand how the acoustic features characterizing vocalizations are perceived by listeners. The previous section reviewed vocal production mechanisms, establishing that oscillation of vocal folds results in sound

spectra characterized by 1) harmonicity: frequency components that are integer multiples of an F0, and 2) spectral modulations: a spectral profile with regular variations in energy across frequencies. Note that harmonic sounds are by default spectrally modulated, but spectrally modulated sounds are not necessarily harmonic, since their energy could be concentrated at frequencies that are not harmonically related. Nonetheless, in vocalizations and other natural sounds generated by periodic vibrations, harmonicity and spectral modulation tend to co-occur. Here, we review the current literature surrounding how humans (Section 1.3) and other animals (Section 1.4) perceive harmonic and spectrally modulated sounds, two typical features of natural vocalizations.

1.3.1 Perception of pitch in harmonic sounds

An exploration of harmonic sound perception would not be complete without discussing pitch perception. In human listeners, harmonic sounds are perceived as a single fused sound with pitch corresponding to the F0, instead of a combination of many different frequencies (Bendor & Wang, 2005). Importantly, the lowest-frequency component (corresponding to F0) need not be present to evoke the perception of a pitch corresponding to F0. This perceptual phenomenon, termed the "pitch of the missing fundamental", suggests that the perception of F0 from a harmonic sound is not merely due to the detection of the lowest-frequency component (Plack & Oxenham, 2005). In the spectral domain, F0 is equal to the highest common denominator of the frequencies composing the spectrum. In the temporal domain, F0 is equal to the inverse of the period of repetition of the sound waveform. Spectral and temporal mechanisms have been proposed for how the auditory system extracts F0.

According to the spectral theory of F0 extraction, an incoming sound is matched to internal harmonic templates corresponding to a range of F0s, and the best match harmonic template determines the perceived F0 of the sound. This mechanism requires the presence of resolved

harmonics – adjacent frequency components that fall into separate auditory filters in the cochlea at the auditory periphery. Auditory filters generally become broader at higher frequencies; in humans only the lowest 5-10 frequency components in a harmonic sound are generally resolved (Plack & Oxenham, 2005; Song et al., 2016).

According to the temporal theory of F0 extraction, unresolved harmonics – adjacent frequency components of a harmonic sound that fall within the same auditory filter in the cochlea – interact at the peripheral auditory system to generate a temporal envelop with periodicity equal to the F0 of the sound. The periodicity is detected by the auditory system to determine F0 (Song et al., 2016).

Pitch strength, defined as the inverse of the smallest change in F0 that can be detected by human listeners, depends on the presence of resolved harmonics. A previous study showed that pitch strength decreased as spectral components were shifted from resolved to unresolved. However, even when all components were unresolved, a weaker pitch could still be perceived (Houtsma & Smurzynski, 1990). Therefore, both spectral and temporal extraction of F0 are thought to play a role in human pitch perception.

1.3.2 Perception of spectral modulation depth in broadband sounds

Spectral modulations, which refer to variations in energy along the frequency axis, are present in speech sounds and may provide cues for perception. Low-rate spectral modulations result from emphasized frequency ranges, known as formants, that are a consequence of filtering by the vocal tract. Higher-rate modulations result from harmonic structure, generated by vibration of the vocal cords. In this dissertation, we mainly focus on spectral modulations at higher rates that are introduced by harmonic structure.

The predominant way in which spectral modulation perception in humans has been assessed is by measuring the minimum spectral modulation depth (spectral peak-to-valley distance in dB) required for listeners to detect a spectrally modulated sound from a flat spectrum sound. Spectrally modulated sounds used in these experiments usually consist of a noise carrier modulated by a sinusoidal spectral envelop. The amplitude of the sinusoidal envelop specifies modulation depth, and the frequency specifies the modulation density. When spectral modulation depth detection thresholds were measured over a range of modulation densities, it was found that listeners' modulation depth detection was the finest at modulation densities of 2 to 4 cycles per octave. At the optimum modulation densities, listeners could detect spectral modulations as low as 2.5 dB (Eddins & Bero, 2007).

Spectral modulation detection may play a role in speech recognition. In a study that measured both spectral modulation detection and speech recognition in the same subjects, modulation detection thresholds at 2 cycles per octave was a significant predictor of speech recognition performance (Davies-venn et al., 2015). In addition, spectral modulation depth detection was improved through practice, but the benefit of practice was only apparent at specific modulation densities. The effect of practice on detection did not generalize across modulation densities nor across frequency ranges (Sabin et al., 2012).

1.3.3 Effect of spectral degradation on speech perception

Many studies in human speech perception utilize spectral degradation to reduce the spectral information of speech sounds into a small number of frequency channels. This method is termed vocoding. Within each channel, signals can be reproduced with a variety of carriers, the most common ones being sine waves centered at each channel and noise bands matching the frequency

range of the channel. These methods are often aimed at mimicking hearing from cochlear implants, which provide impoverished spectral information delivered through a limited number of channels. Spectral degradation of harmonic sounds using a noise vocoder causes adjacent frequency components to “blend in” with one another when the number of channels is low. This concurrently diminishes the harmonic structures and deep spectral modulations characterizing natural vocalizations.

Speech recognition was found to be remarkably robust to spectral degradation. Reducing speech to only three spectral channels, consisting of temporally modulated noise bands, still permitted near-perfect identification of vowels, consonants, and words. This finding illustrates that speech can be understood in the absence of harmonic structure and spectral modulations (Shannon et al., 1995).

The perception of speaker characteristics, however, is more sensitive to spectral degradation. One study found that at least 16 spectral channels in a sine-wave vocoder were required for listeners to discriminate between female and male speakers (Fu et al., 2005). Another study found that the accuracy of speaker sex discrimination from noise-vocoded speech increased when the number of channels was increased from 4 to 10. Listeners’ ability to identify the speaker also increased with the number of channels. Speaker sex discrimination and speaker identification with sine-wave vocoded signals were shown to be less sensitive to number of channels than with noise-vocoded signals (Gonzalez & Oliver, 2005).

In a related study, a modulation filtering technique was used to selectively degrade spectral modulations of particular densities in speech. Speech comprehension and voice sex discrimination were found to be affected by spectral degradation at different resolutions (Elliott & Theunissen, 2009). In this study, sentences embedded in Gaussian white noise were presented, and listeners

were asked to type all the words that they could hear, and whether they perceived the speaker to be male or female. While speech comprehension was impaired when low density spectral modulations were removed (< 4 cycles/kHz), speaker sex discrimination was impaired by removal of higher modulation densities (3 - 7 cycles/kHz) (Elliott & Theunissen, 2009).

Taken together, the current literature suggests that different aspects of speech communication rely on different vocalization-typical spectral features. While harmonic structure and fine spectral modulations are not necessary for speech intelligibility, they contribute significantly to the extraction of speaker characteristics (sex and identity) from vocal quality.

1.3.4 Importance of spectral features of speech in the cocktail party problem

In real-world listening, we often encounter sounds from different sources that overlap in frequency and time. One important task of listeners in complex auditory environments is to parse and track sounds coming from different sources, despite spectral and temporal overlap. This challenge is also referred to as the "cocktail party problem."

Harmonicity is thought to be an important cue in grouping sounds from a common source. For example, perceptual grouping of harmonically related frequencies could aid in separating speech sounds from two speakers whose voices have different F0s. A recent study investigated how harmonicity contributes to sound source segregation (Popham et al., 2018). In this study, human listeners were presented with two concurrent words. The words were either harmonic or inharmonic, with frequencies components randomly shifted up or down. The intelligibility of concurrent words was lower for inharmonic speech compared to harmonic speech. Inharmonic speech was also more difficult to identify and track over time. Listeners were presented with concurrent sentences uttered by different speakers, and asked to focus on one target speaker and report the last word spoken. The accuracy of word reporting was lower for inharmonic speech

compared to harmonic speech, and listeners were more likely to report a word from the non-target speaker when speech was inharmonic.

In addition to investigating the effects of inharmonicity on sound segregation, the authors of this study also studied noise-excited speech, where the discrete frequency components in speech were replaced by noise. In addition to removing harmonicity, this manipulation also removed the spectral modulations resulting from distinct frequency components composing speech. Noise excitation further reduced concurrent word intelligibility beyond the effect of inharmonicity (Popham et al., 2018). These results indicate that while concurrent voice segregation depends on harmonic frequency relationships, the presence of spectral modulations may play an additional role.

1.4 PERCEPTUAL AND BEHAVIORAL IMPORTANCE OF VOCALIZATION-TYPICAL SPECTRAL FEATURES IN NON-HUMAN ANIMALS

1.4.1 Perception of pitch in harmonic sounds

Some of the strongest parallels between pitch perception in humans and that of a nonhuman animal have been described in studies of the common marmoset *(Callithrix jacchus).* The marmoset is a New World primate that has emerged as a prominent model for studying auditory perception and coding. Robust vocal communication behavior in the laboratory setting and recent development of molecular and neurophysiological techniques allow for rich experimental opportunities (Eliades & Miller, 2016).

A recent study has shown that marmosets perceive pitch from harmonic complex sounds in a similar way as do humans (Song et al., 2016). Marmosets, like humans (Section 1.2.1), rely more on resolved harmonics than unresolved harmonics to perceive pitch; the lower limit for F0 difference detection was smaller for stimuli with resolved harmonics than for stimuli with

unresolved harmonics. When discriminating F0 of harmonic sounds with only resolved harmonics, marmosets are sensitive to the integer frequency ratios between frequency components. When frequency components were shifted up or down to render the sound inharmonic, the ability to discriminate F0 differences decreased. When discriminating F0 of harmonic sounds with only unresolved harmonics, marmosets were sensitive to the temporal structure of the waveform. When the phase of individual frequency components was varied such that the composite waveform of the sound is flattened, marmosets had greater difficulty in discriminating F0. Taken together, these studies show that marmosets likely utilize both spectral and temporal cues of harmonic sounds to extract F0.

Some species of non-human animals have been shown to perceive the pitch of the missing fundamental, similar to that described in humans. In these studies, animals were typically trained to perform certain behavioral tasks that required associating certain behaviors with either pure tones (where frequency equals F0), or "missing fundamental" harmonic sounds (MF harmonics; these sounds lack the lowest frequency component, but contain higher frequency components that are integer multiples of the F0). After training, animals are required to generalize behavioral associations either from pure tones to MF harmonics, or from MF harmonics to pure tones with the same F0. Since pure tones and MF harmonics do not have overlapping spectral components, generalization of behavior between the two types of stimuli are taken to indicate that animals extracted F0 from these sounds to perform the behavioral task. Studies of the missing fundamental percept in songbirds and cats are described below.

In a study of missing fundamental percept in European starlings *(Sturnus vulgaris)*, birds were trained to discriminate between two pure tones with different frequencies. When they were tested with two MF harmonics with F0 that matched the frequency of the training tones, the

discrimination persisted. Similar behavioral results were obtained when starlings were trained with MF harmonics and tested with pure tones (Cynx & Shapiro, 1986).

Cats were also shown to perceive the missing fundamental in a study where they were trained to either lick or not lick a drinking spout based on the frequencies of a pair of pure tones, where the second one was either lower or higher in frequency than the first one. Cats were tested with MF harmonics, in which the direction of absolution frequency changes in its spectral components were the opposite of the direction of F0 change. When tested with these stimuli, cats behaved according to the direction of F0 changes, indicating that cats perceived the F0 of these sounds instead of the frequency of individual components (Heffner & Whitfield, 1976).

1.4.2 Perception of spectral modulation depth in broadband sounds

Studies of spectral modulation depth detection in non-human animals are relatively scarce compared to those in humans. Here we describe two studies that have probed the detection of spectral modulation in avian species, budgerigars *(Melopsittacus undulatus)* and zebra finches.

Spectral modulation detection was tested in budgerigars using broadband rippled sounds with sinusoidal spectral envelops. The minimum modulation depths allowing birds to detect rippled sounds from flat-spectrum noise were measured as a function of the spectral modulation density. At low spectral modulation densities (0 - 4 cycles per octave), birds' discrimination thresholds ranged between 2 to 3 dB peak-to-valley modulation depth, whereas human thresholds obtained with the same procedure ranged from 3 to 4 dB. Though modulation depth sensitivity was similar between budgerigars and humans at low modulation densities, at higher modulation densities, budgerigars were more sensitive to modulation depth than human listeners (Amagai et al., 1999).

A more recent study in both budgerigars and zebra finches also tested modulation depth discrimination with spectrally rippled sounds (Osmanski et al., 2009). Unlike the previous study described, which used rippled sounds whose spectral envelops were static over time, this study utilized rippled sounds with sinusoidal spectral envelops that moved up or down in frequency at a constant rate. Budgerigars, zebra finches, and humans showed similar modulation depth discrimination thresholds, and all three species' discrimination ability worsened with increasing modulation density and increasing rate of frequency movement. Interestingly, both species of birds were more sensitive in discriminating rippled sounds that moved down in frequency than those that moved up in frequency. It was postulated that this directional preference was due to the greater prevalence of downward frequency modulations in these birds' natural vocalizations.

While budgerigars and zebra finches show spectral modulation depth sensitivities greater than or comparable to humans, macaques (*Macaca mullata*) have found to show worse spectral modulation depth discrimination ability than humans: modulation depth detection thresholds were 12-20 dB in macaques, compared to 3.5-7.4 dB in humans (O'Connor et al., 2000).

1.4.3 Importance of spectral features in sound recognition and categorization

Human studies have extensively tested the effect of spectral degradation on communication tasks including speech recognition and speaker identification. In non-human animals, the effect of spectral degradation on communication-related behavior is less well understood. However, recent studies have investigated the effect of spectral degradation on animals' ability to navigate the auditory world. Below we review two examples of recent studies that describe how songbirds utilize spectral cues to recognize sound sequences and classify sounds into distinct categories.

One recent study identified the spectral cues that starlings use to recognize sound sequences (Bregman et al., 2016). In this study, starlings were trained to recognize sequences of four

harmonic sounds that either increased or decreased in frequency. Each element in the sequence was a computer-generated sound corresponding to a different musical instrument, such that spectral shape, in addition to F0, differed between elements. Birds were trained to discriminate between sound sequences with ascending and descending F0, then tested with spectrally degraded sound sequences. When tested with vocoded versions of the training sequences, birds were still able to perform the discrimination. It was thus concluded that starlings relied on coarse spectral shape, which was preserved in vocoded versions of sound sequences, to perform sound sequence recognition. Hence, harmonic structure and associated spectral modulations, which were disrupted by the vocoding manipulation, were not required for starlings to recognize sound sequences.

Another study conducted in zebra finches examined the categorization of artificial vowel sounds by either F0 or spectral shape (Burgering et al., 2018). Birds were trained to classify six harmonic vowel sounds into two categories, either based on their F0, or based on their spectral shape, which depended on the relative amplitude of frequency components. Birds that were trained to categorize based on spectral shape were able to categorize novel vocoded sounds based on spectral shape and ignore the absence of harmonic structure. Birds that were trained to categorize vowel sounds based on F0 were able to generalize to novel harmonic sounds and ignore spectral shape. Taken together, this shows that zebra finches can categorize sounds using either F0 or spectral shape. When spectral shape was the relevant parameter, birds' categorization was robust to spectral degradation. Hence, the dependency of sound categorization on spectral structure is flexible and dependent on behavioral context.

1.5 NEURAL PROCESSING OF VOCALIZATION-TYPICAL SPECTRAL FEATURES

1.5.1 Enhanced representation of harmonic sounds in the auditory cortex

The observation that harmonic structure is abundant in natural vocalizations gives rise to the question of whether they are represented and processed by specialized neural populations in the auditory system. The strongest evidence to date for neural sensitivity to harmonic structures have been described in non-human primates. However, it is unknown whether the neurons representing harmonic structure play a role in vocalization processing and in animals' perception of harmonic sounds.

Neurons that may represent harmonic templates have been identified in the marmoset primary auditory cortex (Feng & Wang, 2017). These neurons' responses were facilitated by combinations of harmonically related tones, and could not be predicted by responses to pure tones alone. Harmonic template neurons were sensitive to perturbations of equal spacing between spectral components of a harmonic; their responses decreased with increasing amounts of frequency perturbation. Harmonic template neurons were spatially mingled with other neurons in the primary auditory cortex, and their preferred frequencies spanned the entire hearing range of marmosets. It is possible that harmonic template neurons serve to recognize harmonic sound patterns in incoming sounds.

Beyond neurons that recognize harmonic templates, another population of neurons in the marmoset auditory cortex have been found to encode F0 of harmonic sounds ("pitch neurons") (Bendor & Wang, 2005). These neurons provided the first demonstration of a neural correlate of the "pitch of the missing fundamental" phenomenon; they responded to pure tones with a particular frequency (corresponding to its F0), and also to MF harmonics with the same F0 but that do not

contain the lowest frequency component. In contrast to harmonic template neurons, pitch neurons were spatially clustered; they were located near the anterolateral border of the primary auditory cortex, in the low-frequency-tuned areas of the tonotopic map.

"Pitch-sensitive" neural structures have also been identified in the human auditory cortex using functional magnetic resonance imaging (fMRI). Several studies have converged on the idea that these neural populations reside at the border of primary auditory cortex, extending out to non-primary auditory cortex (Lewis et al., 2009; Norman-Haignere et al., 2013; Norman-haignere et al., 2016; Penagos et al., 2004; Puschmann et al., 2010). However, the definitions of "pitch-sensitive" regions in human studies have tended to be less restrictive than those used in neurophysiological studies in non-human primates. For example, in one recent study, pitch-sensitive brain areas were defined as those that showed a greater response to synthetic harmonic stacks than to noise (Norman-Haignere et al., 2013). Hence, while pitch neurons in Bendor and Wang (2005) refer to neurons tuned to a *specific* F0, pitch-sensitive regions in this study refer to neural structures that respond to periodic sounds that *have* F0 (i.e. those that are known to evoke the perception of pitch). In this study, "pitch-sensitive" brain areas responded more to sounds with resolved harmonics (which is known to elicit a stronger pitch percept) than sounds with unresolved harmonics (which evoke weaker pitch). "Pitch-sensitive" responses were localized to a specific area within the auditory cortex, extending from the low-frequency-tuned regions of primary auditory cortex (anterolateral Heschl's gyrus) anteriorly to non-tonotopic nonprimary auditory cortex (Norman-Haignere et al., 2013). The location of "pitch-sensitive" brain areas may correspond to the location of pitch neurons found in Bendor and Wang (2005), potentially indicating a conserved center for harmonic pitch processing in the primate brain.

Another study assessed human auditory cortex fMRI responses to a suite of natural and synthetic sounds that varied in harmonic-to-noise ratio (HNR), which quantifies the strength of periodic energy in a sound relative to aperiodic energy (Lewis et al., 2009). For example, animal screeches and howls have greater HNR than hisses, and human singing has greater HNR than whispered speech. HNR-sensitive brain areas, which respond more to synthetic and natural sounds with increasing HNR, were localized to portions of Heschl's gyrus and medial superior temporal gyrus (mSTG). Because HNR-sensitive regions were situated in between primary auditory cortex and speech-selective areas, it was proposed that the detection of HNR could serve as an intermediate step to extract vocalizations in humans.

1.5.2 Neural sensitivity to spectral modulations

Previous studies in non-human mammals including ferrets and cats have characterized auditory cortex neurons by their responses to parameters of spectral modulation in broadband sounds. Here we focus on studies that examine spectral modulations in the absence of temporal modulations (Schreiner & Calhoun, 1994; Shamma et al., 1995). In these studies, stimuli were broadband sounds with spectral envelops that were sinusoidal along the logarithmic axis. Spectral modulation density is determined by the frequency of the envelop sinusoid, with higher envelop frequency resulting in more closely spaced spectral peaks. Spectral modulation depth is determined by the amplitude of the envelop sinusoid, with greater envelop amplitude resulting in greater contrast between spectral peaks and valleys. Spectral modulation phase is determined by the starting phase of the envelop phase, with a positive shift in phase moving all peaks and valleys to lower frequencies.

In the ferret primary auditory cortex, the majority of neurons were found to be tuned to specific modulation densities, indicating that different neurons may specialize in detection of

spectral structure at different resolutions (Shamma et al., 1995). Neural responses were also sensitive to ripple phase, showing the highest responses at particular spectral peak placements. The modulation density response function (showing how neural responses varied with modulation density) generally increased in amplitude with increasing modulation depth, but their shape did not change with depth. Neurons in the cat primary auditory cortex were also similarly sensitive to modulation density, depth, and phase (Schreiner & Calhoun, 1994). They exhibited tuning to specific modulation densities, and preferred modulation phases where spectral peaks matched the preferred frequency. When modulation density and phase was held constant, responses increased with modulation depth and plateaued when a maximum depth is reached.

1.6 THE SONGBIRD AS A MODEL FOR VOCAL COMMUNICATION

The study of auditory mechanisms mediating vocal communication necessitates the use of multidimensional approaches in an experimentally tractable system. A critical aspect of human vocal communication is the ability to learn to produce speech during early life. Few other animals, including cetaceans (Janik, 2014), bats (Prat et al., 2015), elephants (Joyce et al., 2005), and songbirds (Brainard & Doupe, 2013), are known to possess the ability to learn to produce complex vocalizations. Songbirds, in particular, have become a prominent system to study the neural mechanisms of vocal learning. Research in songbirds throughout the past decades have uncovered specialized motor and sensory brain regions underlying vocal production and perception (Brainard & Doupe, 2013). Anatomical and electrophysiological studies showing parallel organization between avian and mammalian auditory cortices provide further support for the informative value of songbirds as a model system (Calabrese & Woolley, 2015; Y. Wang et al., 2010).

The zebra finch is a songbird species commonly used to study the mechanisms of vocal communication. Male zebra finches learn to produce complex vocalizations termed song in early

life through a process of hearing, imitation, and practice. Song is typically learned during a closed critical window and crystallizes at adulthood. Learned song is used in courtship and is a critical cue for mate selection by females (Hauber et al., 2010; Zann, 1996). Besides song, both female and male zebra finches have a wide repertoire of vocalizations used in distinct behavioral contexts. Distinct vocalizations categories are used between mates during pair bonding and nest building, between birds to establish contact over short and long distances, and between adults during aggressive encounters or threatening situations (Elie & Theunissen, 2016). Zebra finches readily and consistently exhibit certain communicative behaviors in the laboratory, providing a solid platform to understand how experimental variables contribute to animals' use of vocalizations in an ethologically relevant setting (Vicario et al., 2001; Vignal & Mathevon, 2011).

Much effort has also been dedicated to understanding how songbirds (not limited to zebra finches) perceive complex auditory stimuli. As described in earlier sections, starlings have been shown to perceive the pitch of the missing fundamental (Cynx & Shapiro, 1986), and to recognize complex sound sequences based on spectral shape (Bregman et al., 2016). Zebra finches can detect modulations in the spectral envelop with comparable sensitivity to human listeners (Osmanski et al., 2009), as well as detect very slight mis-tuning in harmonic frequency components with sensitivity beyond that of human listeners (Lohr & Dooling, 1998). Much remains to be learned about how songbirds engage complex auditory perceptual mechanisms in vocal communication and the neural mechanisms governing this process.

1.7 CONCLUSIONS

Vocalization processing is a critical component of acoustic communication. The vocal sounds of many animals are characterized by harmonicity and deep spectral modulations, suggesting that specialized neural mechanisms may be in place to extract these features.

Remarkable advances have been made in understanding how complex spectral features are perceived, and in identifying possible neural substrates of perception. However, many previous studies address perception and neural mechanisms in isolation, without considering the behavioral relevance of the stimuli used. The following two considerations are proposed for further studies of auditory processing in vocal communication.

First, neurophysiological and neuroimaging studies of complex spectral processing have predominantly utilized synthetic sounds to probe neural function. The ability to easily control and manipulate stimulus parameters in synthetic sounds has proven useful in understanding the principles of auditory processing. Studying neural responses to vocalizations or vocalization-like sounds with known behavioral significance will allow a better understanding of auditory mechanisms active in vocal communication. In **Chapters 2** and **3,** I describe our use of acoustic stimuli that capture behaviorally relevant features of vocalizations. These stimuli are used to identify neural populations that could support vocal communication. Second, a common approach has been to contrast responses to harmonic sounds with responses to noise in order to identify brain regions sensitive to pitch and harmonic structure. As harmonic sounds contain deep spectral modulations, selectivity for harmonic sounds over noise could in fact be driven by spectral modulations instead of harmonicity. Dissociating harmonic structure and spectral modulations form the basis of studies described in **Chapter 4.**

Chapter 2

ACOUSTIC, BEHAVIORAL, AND NEUROPHYSIOLOGICAL METHODS TO INVESTIGATE AUDITORY PROCESSING OF VOCALIZATIONS

2.1 INTRODUCTION

The quest to understand how the auditory system functions during social communication necessitates a combination of acoustic, behavioral, and neurophysiological approaches. The choice of acoustic stimuli – the sounds used to investigate neural and behavioral responses, is particularly important. Synthetic stimuli, such as pure tones (e.g. Kikuchi et al., 2014), two-tone combinations (e.g. Shamma et al., 1993), and band-passed noise (e.g. Rauschecker et al., 1995) offer the advantage of permitting easy control over stimulus parameters. On the other hand, natural or naturalistic stimuli, such as vocalizations, are more representative of the challenges that the auditory system encounters and must resolve in real-world situations, and are often more reliable at eliciting responses from auditory neurons, especially at higher levels of the auditory system such as the cortex (Theunissen & Elie, 2014).

In addition to the acoustic structure and complexity of sound stimuli, it is also important to consider what sounds are behaviorally important to animals whose neural systems are being studied, before investigating the neural activity patterns elicited by these sounds. Using stimuli that engage natural communication behaviors lends support to the potential relevance of our studies in real-world listening situations. Three overarching considerations in experimental design

are described below, and details of methodologies are further discussed in the subsequent sections of this chapter.

First, I designed synthetic acoustic stimuli that varied systematically in acoustic features of interest, using natural vocalizations as templates. Noise-vocoded calls, where spectral information in natural vocalizations was reduced into a limited number of channels, were generated and used in behavioral and neural experiments. Inharmonic calls, in which the harmonic relationships between frequency components are disrupted, were used in behavioral experiments. While both noise-vocoded calls and inharmonic calls were synthetic sounds, they elicited social responses from birds. This indicated that these stimuli sufficiently captured the behaviorally relevant qualities of natural vocalizations, making them appropriate for studying auditory processing in the context of social communication. Further, both types of stimulus design have been used in studies of perception in humans and other animals. Noise-vocoded vocalizations have been used to study speech perception in humans (Gonzalez & Oliver, 2005; Shannon et al., 1995), in a behavioral study in European starlings (Bregman et al., 2016), and in a neurophysiological study in gerbils (Ter-mikaelian et al., 2018). Inharmonic versions of speech and music sounds have been generated and used recently for perceptual studies in humans (McDermott et al., 2012; McPherson & McDermott, 2018; Popham et al., 2018), paving the way for investigating the validity of long-standing theories about the role of harmonicity in sound segregation and pitch perception of natural sounds. Parallels between acoustic manipulations between our studies and previous studies allow us to relate our findings to studies in other species and in other behavioral contexts.

Second, I implemented a behavioral paradigm that utilizes a natural behavior, namely, the tendency of birds to produce contact calls in response to other birds' vocalizations (Vicario et al.,

2001), in order to identify important acoustic features driving social responses. This behavioral paradigm does not involve operant conditioning and allowed me to determine what types of sounds are intrinsically behaviorally relevant (i.e. evoking significant responses), and what sounds are not. Then, when presenting the same set of sounds and analyzing the responses of auditory neurons, I was able to identify which neurons in the auditory system can effectively process the acoustic parameters that distinguish behaviorally relevant sounds from irrelevant ones.

Third, I assessed auditory cortical responses to sound stimuli with a range of acoustic complexity. These stimuli included pure tones – sinusoids with a constant frequency; ripples – broadband stimuli whose frequency structure varied systematically; calls (natural and synthetic) – communication signals with known behavioral importance, as assessed from behavioral experiments; and finally, songs – a combination of sound elements with diverse acoustics. Experiments with various stimulus sets, combined with a neurophysiological approach that allowed simultaneous recordings of single neuron responses from 32 sites in the auditory cortex, allowed me to derive fundamental principles that govern neurons' responses across divergent stimuli.

2.2 MANIPULATING SPECTRAL FEATURES OF VOCALIZATIONS

2.2.1 Recording and selecting natural vocalizations

For behavior and electrophysiology experiments, natural distance calls were used as stimuli. Distance calls were recorded from adult female zebra finches and band-pass filtered between 300 and 8000 Hz. For call recordings, females were housed in isolation in a sound-isolated chamber (Industrial Acoustics), with ample access to food and water. Their vocalizations were recorded through a microphone (Sennheiser MKE 2-60) connected to an audio interface (Focusrite Saffire Pro 40) using the recording function in Sound Analysis Pro software (Tchernichovski et

al., 2000). Sounds in the booth were continuously monitored and recording was automatically initiated when the sound amplitude exceeded a user-determined threshold. All audio recordings were subsequently manually sorted to extract distance calls and to separate them from other vocalizations and non-vocalizations sounds resulting from the birds' movement. Nine natural calls from nine different females were included as stimuli in behavioral and neurophysiological experiments and were used as templates to create synthetic calls.

I chose to use female distance calls as acoustic stimuli for all experiments because of their known behavioral function, as well as well-defined and relatively simple acoustic features. Distance calls are classified as long-range contact calls that are used for communication over long distances. They are commonly produced when birds are out of visual range with the colony, their mate, or the offspring that they care for (Elie & Theunissen, 2016). Distance calls contain individual signatures, and zebra finches can recognize their mates using these calls (Vignal et al., 2008).

Compared to other vocalizations in the zebra finch repertoire, distance calls have the strongest harmonic structure. In addition, they tend to be louder and longer in duration than the other class of contact calls (also known as tet calls), which are used for short-range communication (Elie & Theunissen, 2016). Distance calls are sexually dimorphic, with female calls having lower fundamental frequencies, less frequency modulation, and longer duration than male calls. Female calls have been shown to elicit greater vocal responses than male calls from both female and male zebra finches (Vicario et al., 2001). Females calls were thus considered ideal stimuli for behavioral experiments, as they provide a higher ceiling to test for the effect of acoustic manipulations on behavioral responses. Acoustic manipulations will be described in the following sections.

To confirm that the 9 distance calls that we chose (stimulus calls) were generally representative of female distance calls, we analyzed the acoustic features of the chosen calls against a pool of other distance calls from 17 females (reference calls). Nine of these reference calls were collected from females in our lab's colony at Columbia University, while the other 8 were collected from females from other universities' colonies and provided by E.C. Perez. Stimulus calls did not differ from reference calls in all the acoustic features analyzed, including duration, aperiodicity, fundamental frequency, goodness of pitch, entropy, mean frequency, and frequency modulation (two-sample t-tests, all $p > 0.05$).

For electrophysiology experiments, we additionally included five songs recorded from male zebra finches in our colony. Songs are produced by male birds as a courtship signal, and young males learn to sing from adult tutors in early life (Zann, 1996). Song is composed of acoustic elements termed syllables, separated by inter-syllable periods of silence and arranged sequentially into motifs. Song stimuli in my experiments were 1.7 to 2.4 seconds in length and chosen to include a range of syllable acoustics. Song were presented as stimuli to a subset of recorded units (1008 single units out of 1825; 4 out of 6 birds). Song syllables are generally broadband, ranging from harmonically structured to noisy (Fee et al., 1998). This provided an opportunity to test how neurons' responses are modulated by natural variations of spectral structure across syllables (and in some cases within syllables). A previous study in our lab has shown that local acoustic context (i.e. the acoustic elements preceding the sound of interest) can affect how individual sound elements are processed (Schneider & Woolley, 2013). By relating call responses to song responses, we could also assess whether neurons' spectral processing properties are specific to certain acoustic contexts.

2.2.2 Generation of noise-vocoded calls

Noise-vocoded calls were generated from the 9 female natural calls using a vocoder implemented in MATLAB (Gaudrain, 2016). For behavioral experiments, the frequency axis of natural calls was divided into 16, 20, 27, 40, 80 linearly-spaced bands. These channel numbers resulted in channel widths that decreased approximately linearly; channel widths were 481 Hz, 385 Hz, 285 Hz, 183 Hz, and 96 Hz respectively (**Figure 2.1A**). Stimuli for electrophysiology experiments included the same vocoded calls as in behavioral experiments, with the addition of 120-channel stimuli (channel width = 64 Hz).

The noise-vocoding procedure included an analysis step and a synthesis step. In the analysis step, the natural distance call was divided into a certain number of channels, and the lower and upper frequency edges of each channel was determined. Within each channel, the frequency-limited distance call was obtained by bandpass-filtering with 12th order Butterworth filters between the lower and upper frequency edges of the channel. Then, the amplitude envelop of the filtered distance call was extracted by half-wave rectification and smoothed by low-pass filtering at 150 Hz with 4th order Butterworth filters. This filtering, envelop extraction, and smoothing process was repeated for each channel and the resulting amplitude envelops were stored for use in the subsequent synthesis step.

In the synthesis step, bandpass-filtered noise signals were created for each channel using lower and upper frequency edges matching those used in the analysis step. Within each channel, bandpass noise was modulated by the amplitude envelopes extracted from natural signals in the analysis step. Finally, all the amplitude-modulated noise bands across all frequency ranges were combined to form a noise-vocoded call (Fig 2.1B).

This analysis and synthesis procedure resulted in noise-vocoded calls that show conserved coarse spectral structure and time-varying amplitude envelops. The power differences across channels are maintained during the synthesis of vocoded calls from natural calls, resulting in spectral shapes that overlapped between vocoded calls with different numbers of channels (**Figure 2.1C**). The amplitude envelop of vocoded calls were highly correlated to that of natural calls, indicating that vocoded calls retain the natural temporal structure of distance calls (**Figure 2.1E**). The main feature that is altered by changing the number of channels is the fine spectral structure that arises from distinct harmonic components found in the natural calls. At low channel numbers (16-27 channels), adjacent frequency components of the natural call fall within the same channel, causing a spectral blurring effect. At higher channel numbers (40-80 channels), adjacent frequency components are resolved by separate channels and result in the emergence of vocalization-typical spectral structure (**Figure 2.1C-D**).

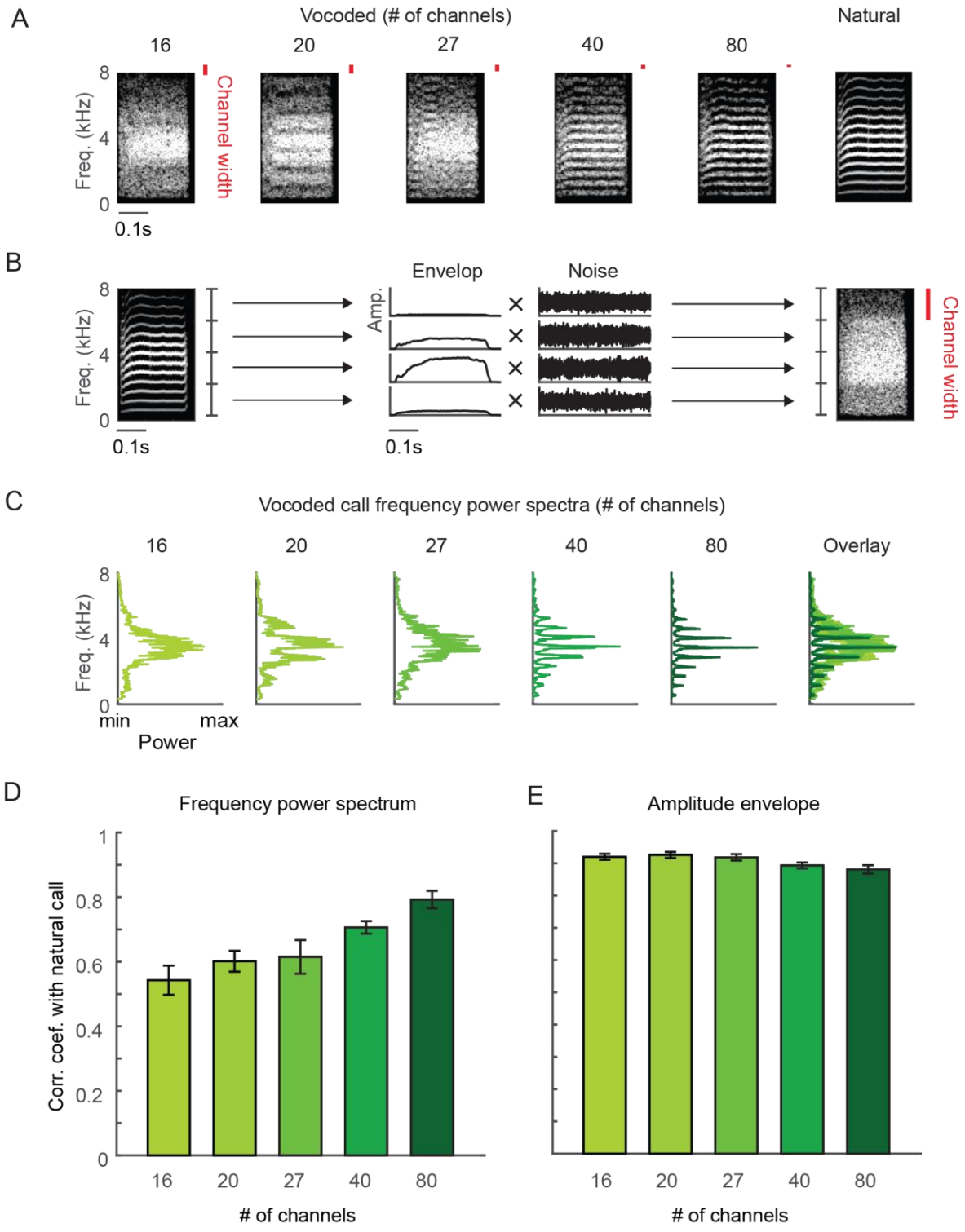


Figure 2.1 Generation of noise-vocoded calls with differing spectral resolution. *(A)* Example spectrograms showing frequency content of vocoded calls and natural call against time. Spectrograms show how spectral structure varies with change in channel number. Red bar indicates the width of one spectral channel. *(B)* Schematic showing how a vocoded call is generated from a natural call. The example depicts generation of a 4-channel vocoded call. The

same process is used to generate vocoded calls with any number of channels. The natural call (left spectrogram) was decomposed into four spectral bands, and the amplitude envelop of each was extracted. Bandpass filtered noise carriers were modulated by the extracted amplitude envelops and recombined to form a vocoded call (right). Red bar indicates the width of one spectral channel. *(C)* Frequency power spectrums (FPSs) of vocoded calls with varying channel numbers. FPSs of vocoded calls with different channel numbers were overlaid and shown in the rightmost graph. While fine spectral structure changed with increases in channel number, coarse spectral shape was maintained. *(D, E)* Correlation coefficient between *(D)* FPSs and *(E)* amplitude envelops of vocoded calls and those of their natural call counterparts (mean ± SEM, N=9). The spectral similarity of vocoded calls with natural calls increased with channel number. Amplitude envelops of vocoded calls were highly correlated with those of natural calls regardless of channel number; correlation decreased only slightly with increases in channel number.

2.2.3 Generation of inharmonic calls

Inharmonic calls were generated in collaboration with M.J. McPherson (Harvard-MIT Program in Speech and Hearing Bioscience) from the same 9 stimulus calls as above, using the modified STRAIGHT framework for speech analysis and synthesis (McDermott et al., 2012; McPherson & McDermott, 2018; Popham et al., 2018). Under the modified STRAIGHT framework, the input signal (a distance call) was decomposed into three time-varying components: spectral envelope, periodic excitation (voiced component), and aperiodic excitation (unvoiced component). The periodic excitation was modelled as a sum of sinusoids, and each sinusoid was then individually modified in frequency. These sinusoidal components were then recombined with the original aperiodic component and time-varying spectral envelop to produce an inharmonic call.

We included synthesized inharmonic calls with varying degrees of inharmonicity, defined by maximum frequency shifts of frequency components. Frequencies of individual components in a distance call were shifted up or down by a random amount (i.e. jittered), and the amount of jitter was constrained within 10%, 30%, or 50% of the F0. For each distance call and maximum jitter amount, we included three variants with different random jitter patters. There was an additional constraint of 30 Hz imposed on the minimum spacing between adjacent frequency components. We also included synthesized harmonic calls, which were generated in the same manner as inharmonic calls, except that frequency shifts were not introduced in the sinusoidal components before synthesis. Synthesized harmonic calls were included to control for any artifacts that may be introduced by the synthesis procedure, which could cause differences between behavioral responses to inharmonic calls and to natural calls, thus confounding our results.

2.3 NON-VOCALIZATION STIMULI FOR AUDITORY NEUROPHYSIOLOGY EXPERIMENTS

2.3.1 Pure tones

In order to characterize the basic frequency response properties of individual neurons, we presented pure tone stimuli in electrophysiology experiments. Pure tone stimuli are sinusoidal signals with a single frequency that does not vary over time. Tones in our stimulus set were 200 ms in duration, including 10 ms linear onset and offset ramps. Pure tone stimuli were generated with frequencies ranging from 500 Hz to 8000 Hz, varying in 500 Hz intervals, and with intensities ranging from 30 dB SPL to 70 dB SPL, varying in 10 dB steps.

2.3.2 Spectrally modulated ripples

Spectrally modulated ripples are broadband sounds consisting of sinusoidal modulations along the frequency axis, and no modulations along the temporal axis. They are considered the auditory equivalent of visual gratings. As described in previous studies (Schreiner & Calhoun, 1994; Shamma et al., 1995), the spectral envelop of ripple stimuli used in our study are determined by three modulation parameters:

1) Density: the frequency of the sinusoidal spectral envelop, which determines how closely spaced the spectral peaks are to one another
2) Depth: the amplitude of the sinusoidal envelop, which determines the peak-to-valley distance in the spectrum
3) Phase: the starting phase of the sinusoidal envelop, which determines the placement of spectral peaks

In our stimulus set, all ripples were 200 ms in duration including10 ms linear onset and offset ramps. All ripples were presented at 60 dB SPL. Ripple stimuli consisted of stacked tones between

250 Hz to 8000 Hz, spaced 1 Hz apart. The relative amplitudes of each tone composed the spectral envelop, which was specified by the spectral modulation density, depth, and phase (**Figure 2.2**).

Our ripple stimulus set included ripples with modulation densities of 1.2, 1.6, and 2.0 cycles/kHz, with modulations applied along a linear frequency scale. We chose to use a linear frequency scale instead of a logarithmic scale, which was used often in previous experiments (Schreiner & Calhoun, 1994; Shamma et al., 1995), because this allows for ripple spectra to be fully aligned with the harmonic spectrum when the maximally-aligning phase is chosen.

The amplitude of sinusoidal modulations was specified on a logarithmic scale, and we included ripples with modulation depths (peak-to-valley distance) of 5 dB, 10 dB, 20 dB, 40 dB, and 80 dB. We also included ripples with 8 different evenly spaced starting phases.

Ripples with different phases had the same spacing between spectral peaks, but the placement of spectral peaks varied. Harmonic ripples were generated when the phase was chosen such that each spectral peak was an integer multiple of a common F0, and the F0 equaled the spacing between peaks. Inharmonic ripples were generated when the phase was chosen such that spectral peaks are shifted with respect to harmonic frequencies. In the case of inharmonic ripples, spectral peaks were still spaced evenly from one another, but the absolute frequencies of each peak were not integer multiples of a common number (F0). Ripples were generated using custom software provided by S. Andoni (University of Texas).

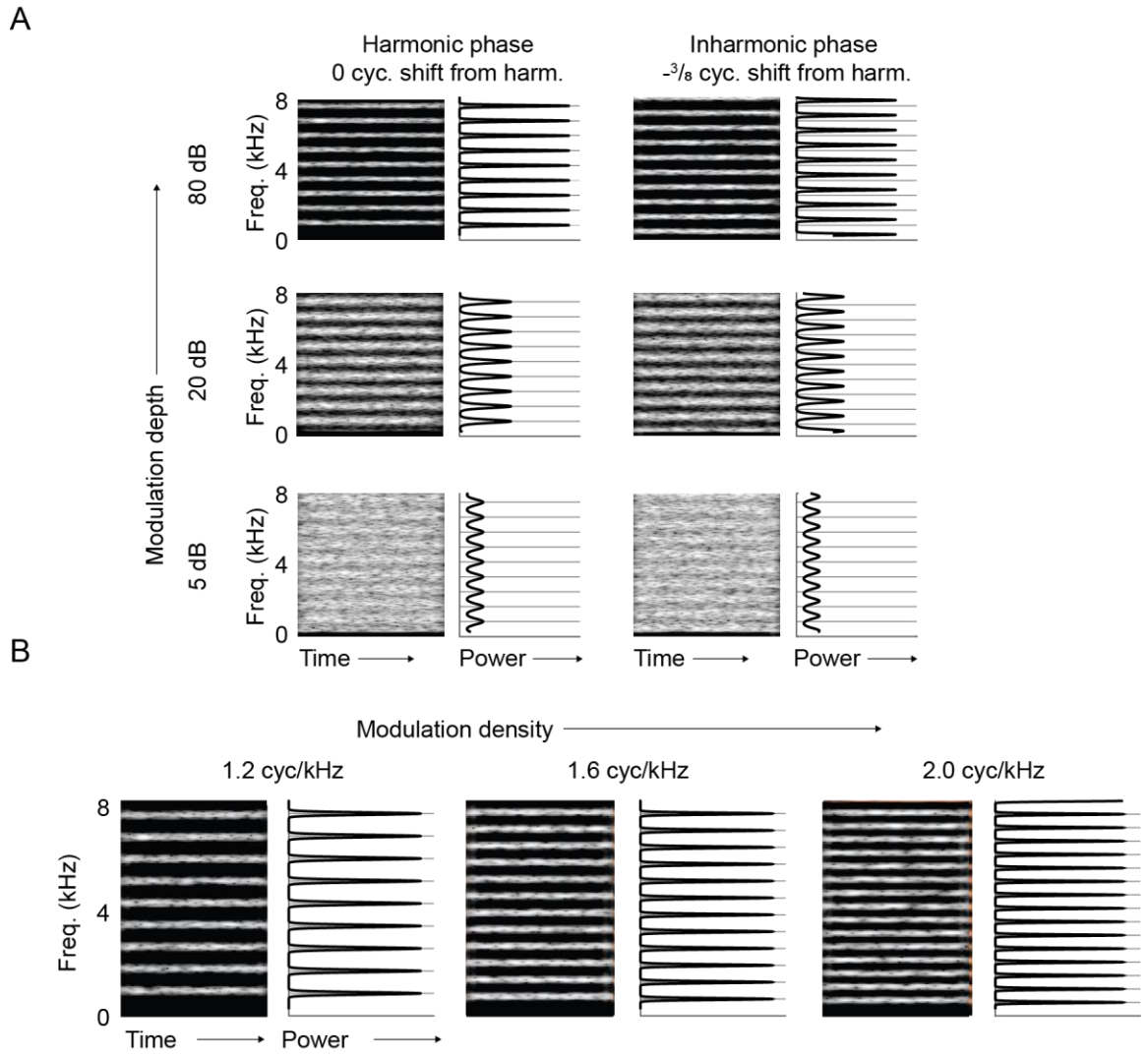


Figure 2.2 Schematic of spectrally modulated ripples with varying depth, phase, and density. *(A)* Spectrograms and spectral profiles of example ripple stimuli that vary in spectral modulation depth and phase. Ripples shown have modulation density of 1.2 cyc/kHz. Spectral profiles are shown to the right of each spectrogram. Depth is varied along the vertical axis. Left column shows ripples with harmonic phase, and right column shows ripples with inharmonic phase. At the harmonic phase, spectral peaks align with integer multiples of a fundamental frequency. Harmonic frequencies (integer multiples of a F0, 833 Hz) are indicated by gray horizontal lines on each spectral profile. *(B)* Spectrograms and spectral profiles of ripple stimuli that vary in spectral modulation density. Shown ripples are of harmonic phase and have modulation depth of 80 dB. Spectral modulation densities are 1.2 cyc/kHz, 1.6 cyc/kHz, and 2.0 cyc/kHz from left to right.

2.4 TESTING BEHAVIORAL RELEVANCE OF VOCALIZATIONS AND VOCALIZATION-LIKE STIMULI

2.4.1 Animals

Vocal responses to call playback were tested in adult male zebra finches (>120 days old; N = 14 for experiment 1, N = 12 for experiment 2). Across all tested birds, 3 were used in both experiments 1 and 2. Prior to testing, birds were isolated for three to five days in an anechoic sound-attenuation booth (Industrial Acoustics) with free access to food and water.

2.4.2 Call-and-response behavioral testing

Experimental sessions for all birds began within three hours following the onset of the light phase of the light/dark cycle and lasted approximately 80 minutes (behavioral experiment 1: vocoded calls) or 90 minutes (behavioral experiment 2: inharmonic calls). For both experiments, stimulus playback and audio recording were controlled by a custom MATLAB program. A maximum of four birds were tested simultaneously in individual sound-attenuation booths. Stimuli sampled at 44.1 kHz were delivered in the free field at 60 dB SPL through a speaker (Kenwood KFC-1377) placed ~24 cm away from the perch in the experimental cage. Birds' vocalizations were recorded using a microphone (Sennheiser MKE 2-60) connected to an audio interface (Focusrite Saffire Pro 40).

2.4.3 Stimulus selection

Behavioral experiment 1: vocoded calls. Each birds' stimulus set contained four natural calls (a subset of the 9 stimulus calls; varied across birds), five vocoded versions of each natural call (with 16, 20, 27, 40, and 80 channels), and a white noise sample that matched the average duration of the four natural calls. Ten repetitions of each unique stimulus were presented in

pseudorandom order. Inter-stimulus intervals were sampled from a uniform distribution between 15 s and 22 s.

Behavioral experiment 2: inharmonic calls. For inharmonic call experiments, each birds' stimulus set contained three natural calls (a subset of the 9 stimulus calls; varied across birds). Also included were nine synthesized inharmonic versions of each natural call. Maximum jitter values were 10%, 30%, and 50%, and three different jitter patterns were included for each maximum jitter value. We also included three synthesized harmonic versions of each natural call, and a white noise sample that matched the average duration of the three natural calls. Eight repetitions of each unique stimulus were presented in pseudorandom order. Inter-stimulus intervals were sampled from a uniform distribution between 15 s and 22 s.

2.4.4 Extracting vocal responses from audio recordings

For both experiments 1 and 2, with each onset of stimulus presentation, audio recording was initiated by a custom MATLAB program to record vocal responses of the subject bird to presented stimuli. Following behavioral testing, audio recordings of the first 10 seconds following stimulus onsets were pre-processed to extract birds' vocalizations using a custom graphical user interface in MATLAB (**Figure 2.3**). First, each audio sample was filtered to remove low-frequency ambient noise. The absolute values of the waveforms were taken, smoothed with a 10 ms moving average, and normalized to have maximum and minimum values of 1 and 0. Second, an amplitude threshold was manually chosen that best separates the birds' vocalizations from stimulus playback and any background noise that was present. Continuous bouts of sound above the manually chosen threshold that corresponded to vocalizations were then selected (**Figure 2.3C**). The onset and offset time of each detected vocalization was then recorded and used for further analysis.

Consistent with previous reports of zebra finches' behavior inside and outside of playback experiments, we found that birds sometimes produced more than one type of call during the experiment, including the louder and longer distance calls, and softer and shorter calls that may represent tet calls (Elie & Theunissen, 2016; Vicario et al., 2001). To distinguish between distinct call types, we extracted each vocalization using their detected time stamps and subjected them to acoustic analysis. For each bird, we isolated distance calls from shorter calls on the basis of duration and mean frequency using k-means clustering on all extracted vocalizations. Details of this analysis are presented in Chapter 3.

2.4.5 Quantification of vocal responses

Distance call responses were quantified using the following two measures of behavior:

1) **Average number of distance calls**: the number of distance calls recorded within the first 5 seconds, averaged across trials
2) **Proportion of trials with response(s):** the number of trials where at least one distance call was emitted in the first 5 seconds, divided by the total number of trials

During pilot experiments we observed that some birds produced weak responses overall, which did not permit reasonable comparison of response strengths to different stimulus types. Because of this, we defined criteria to determine which birds to include in further data analysis. In both Experiments 1 and 2, only birds whose behavioral responses met both of the following criteria were included in our dataset:

1) Produced at least one distance call in the first 5 seconds in ≥10% of all trials, across stimulus types
2) Produced at least one distance call in the first 5 seconds in ≥50% of all trials for at least one specific stimulus

By these criteria, 10 out of 14 birds were included in Experiment 1, and 8 out of 12 birds were included in Experiment 2.

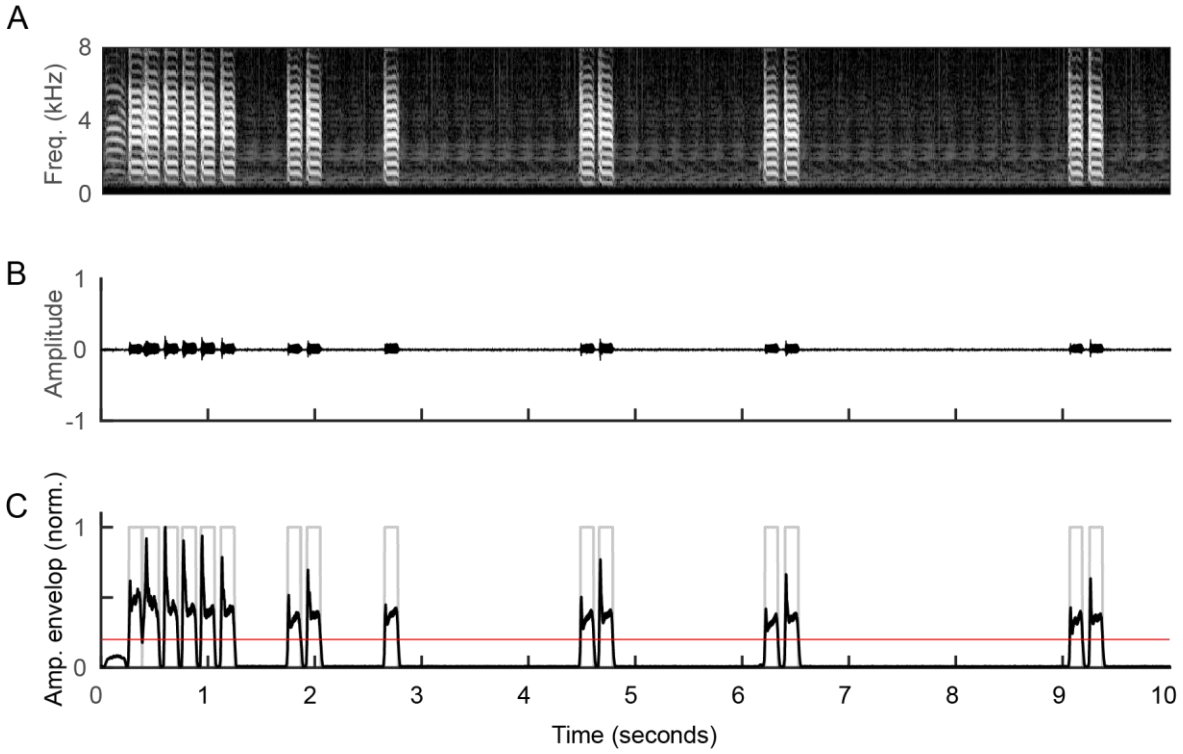


Figure 2.3 Extracting response calls in the call-and-response experiment.
(A) Example spectrogram of a 10 s recording collected from a single trial of the call-and-response experiment. The audio recording includes the call playback, followed by 15 vocalizations emitted by the bird being tested. *(B)* Sound pressure waveform of the same audio recording shown in *(A)*. *(C)* Smoothed absolute value of the sound pressure waveform, normalized to range between 0 (minimum) and 1 (maximum). The red horizontal line indicates a user-defined amplitude threshold applied to extract vocalizations. Gray rectangular outlines indicate continuous bouts of sound with amplitude above the user-defined threshold.

2.5 RECORDING AND ANALYZING NEURAL ACTIVITY IN AUDITORY CORTEX

2.5.1 Animals

Recordings were conducted in 6 adult male zebra finches (>120 days old); 5 out of 6 were used in call-and-response experiments prior to electrophysiology experiments.

2.5.2 Surgery

For surgeries, anesthesia was induced with 2% isoflurane and maintained with 0.5-1.5% isoflurane delivered in 100% oxygen. For induction, isoflurane was delivered through gas tubing connected to an outlet composed of a 50 mL conical tube, which was placed over the bird's head. After birds were unresponsive to a toe pinch, they were enclosed in a custom jacket and placed in a stereotaxic holder. A feeding needle inserted into the bird's beak was used to deliver 0.5 - 1.5% isoflurane for maintenance of anesthesia.

An incision was made on the skin overlaying the skull, and the outer layer of the skull was removed as needed to reveal anatomical landmarks. The bifurcation of the midsagittal sinus was used as the reference point for measuring anatomical coordinates. Then, 2.5 mm by 2.5mm bilateral craniotomies were made, centered at 1.25 mm lateral and 1.25 mm from the reference point on each hemisphere. Using dental acrylic, a metal pin was attached to the skull directly behind the craniotomies. A silver ground wire was inserted beneath the skull and affixed with dental acrylic at a position ~0.2 mm caudal to the bifurcation of the midsagittal sinus. After surgeries, birds recovered for 2 days before the first recording session. Before the first recording session and between sessions, craniotomies were covered with Kwik-Cast Sealant (World Precision Instruments).

2.5.3 Auditory electrophysiology

Neurophysiological recordings were conducted in a walk-in sound-attenuating booth (Industrial Acoustics) in non-anesthetized, head-fixed animals. Prior to each penetration, electrode arrays were coated with CM-DiI (C7000, Molecular Probes) or SP-DiO (D7778, Molecular Probes) dissolved in 100% ethanol. We alternated the use of DiI and DiO between adjacent passes along the medial-lateral axis such that they can be resolved in subsequent histological analysis. One to two electrode penetrations were made per day, with probes oriented along the rostral-caudal axis. Probes were composed of 4 shanks with 8 recording contacts on each shank (Neuronexus A32; **Figure 2.4B**). The spacing was 200 μm between shanks and 100 μm between contacts on the same shank.

Acoustic stimuli were sampled at 24.4 kHz and delivered through a speaker (JBL Control I) placed 23 cm in front of the bird. All non-tone stimuli including natural calls, vocoded calls, ripples, and songs were delivered at 60 dB SPL. During recording sessions, 10 repetitions of each stimulus were presented in pseudorandom order, with inter-stimulus intervals sampled from a uniform distribution spanning 0.75 s and 1 s. Neural responses were recorded at three or four depths along the same pass, with the base of the probe positioned at 1.2 mm, 2.0 mm, and 2.8 mm, or 0.8 mm, 1.6 mm, 2.4 mm, and 3.2 mm below the surface of the brain. Continuous voltage traces were amplified, bandpass filtered between 300 and 5000 Hz, digitized at 24.4kHz (RZ5, Tucker-Davis Technologies), and stored for subsequent data analysis.

2.5.4 Histology and construction of neural maps

After the last recording session, a bird was given an overdose of Euthasol and transcardially perfused with saline followed by 10% formalin. The brain was extracted, separated into the left and right hemispheres, and post-fixed in 10% formalin. After at least 24 hours, the brain was

transferred to 30% sucrose/10% formalin solution for cryoprotection. After cryoprotection, 40 µm parasagittal sections were made using a freezing microtome.

Microtome sections were mounted and, while wet, imaged under CY3 and FITC filters to localize fluorescent DiI and DiO tracks. A bright field image was additionally taken for each brain section as the thalamorecipient region L2a can be distinctly identified from these images as an area of dark fibers (**Figure 2.4A**). After obtaining fluorescent and bright-field images, the sections were dried, stained with cresyl violet (Nissl staining), and imaged to delineate cortical regions by examining cytoarchitectural features.

By visualizing the laminae, cellular appearance (Nissl stain) and dark fibers (from bright field images taken prior to Nissl staining), we determined the boundaries between regions (**Figure 2.4B**). The intermediate region L2a (intermediate-a) is characterized by densely packed cells with small, oblong cell bodies in Nissl images, and by the termination of dark thalamic fibers in wet bright-field images. The intermediate region L2b (intermediate-b) is a population of densely packed, darkly Nissl-stained cells dorsal to the tip of intermediate-a. Deep region L3 is ventral to the border of intermediate-a, and is surrounded dorsally by intermediate-b and caudally by L, which is characterized by dark, densely packed cells. Cells in the deep region are larger and less densely packed than intermediate-a, intermediate-b, and L. The cells are organized into clusters, giving the region a punctate appearance. The ventral border of the deep region is the dorsal medullary lamina (LMD). The secondary region NC is the area posterior to intermediate-b and L.

To create anatomical maps of single unit SSI in the auditory cortex, we estimated the coordinates of each unit by measuring the location of recording sites relative to anatomical reference points that we defined. Reference points were anatomical landmarks that could be

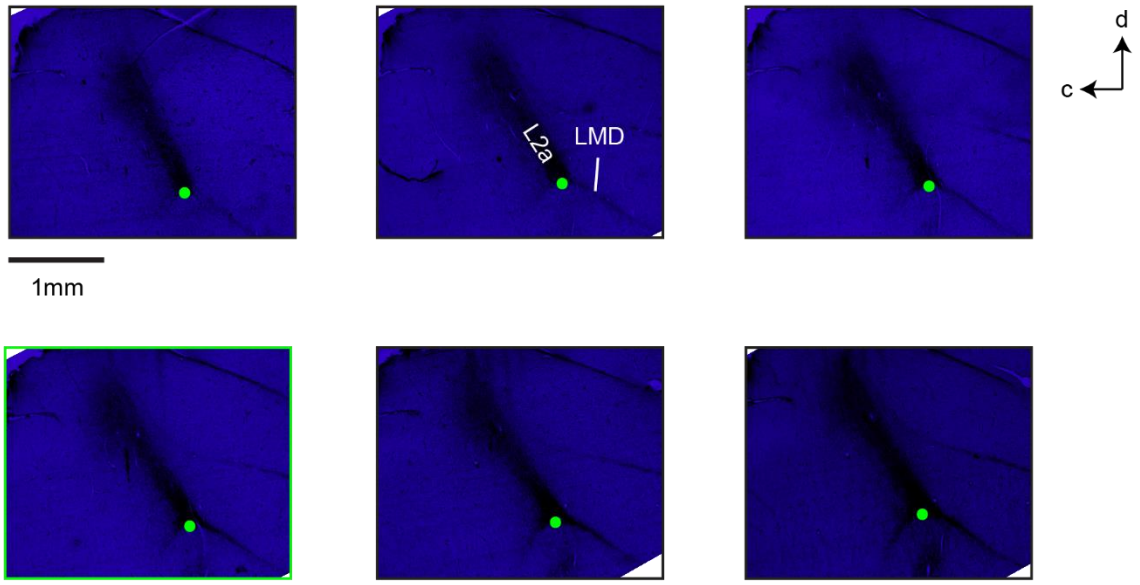
identified in each hemisphere for each bird, serving to standardize recording location estimates across birds.

To estimate the medial-lateral coordinates of all units recorded from a single electrode penetration, we determined a reference plane, which is the parasagittal section at which the ventral tip of L2a fully intersects with and fans out at the LMD when moving laterally from the midline (**Figure 2.4A**). For each identified DiI or DiO probe track, the medial-lateral coordinate was determined based on the relative position of the DiI or DiO track compared to the reference plane. When DiI or DiO signals spanned multiple sections (either due to the spread of dye across sections, or due to the orientation of the probe not being exactly aligned with the plane of sectioning), the most medial and most lateral section containing fluorescent signal was determined, and the section in the middle of those two endpoints was used to determine the medial-lateral coordinate of the units recorded from the corresponding penetration.

To estimate the rostral-caudal and dorsal-ventral positions within a parasagittal section, the reference point was defined as the point at which the ventral tip of L2a is closest to, or intersects with, the LMD. Identified DiI and DiO tracks were used to reconstruct the anterior-posterior and dorsal-ventral coordinates relative to the reference point. The position of the center of the probe base (center-base) was measured relative to the reference point, and a coordinate was assigned to each unit based on the known positional difference between the site containing the units' signal and the center-base of the probe. When necessary, a heatmap was constructed with the standard deviation (SD) of the spontaneous multi-unit activity (MUA) or the stimulus-driven MUA of each channel (**Figure 2.4B**), arranged according to the spatial positioning of different channels. Higher standard deviation and stimulus-driven MUA firing rates characterize L2a and L2b recording sites,

giving them a distinct appearance on the heatmaps. These heatmaps were used to adjust the depth estimate of the probe center-base.

A


L2a
LMD
d
c
1mm

B

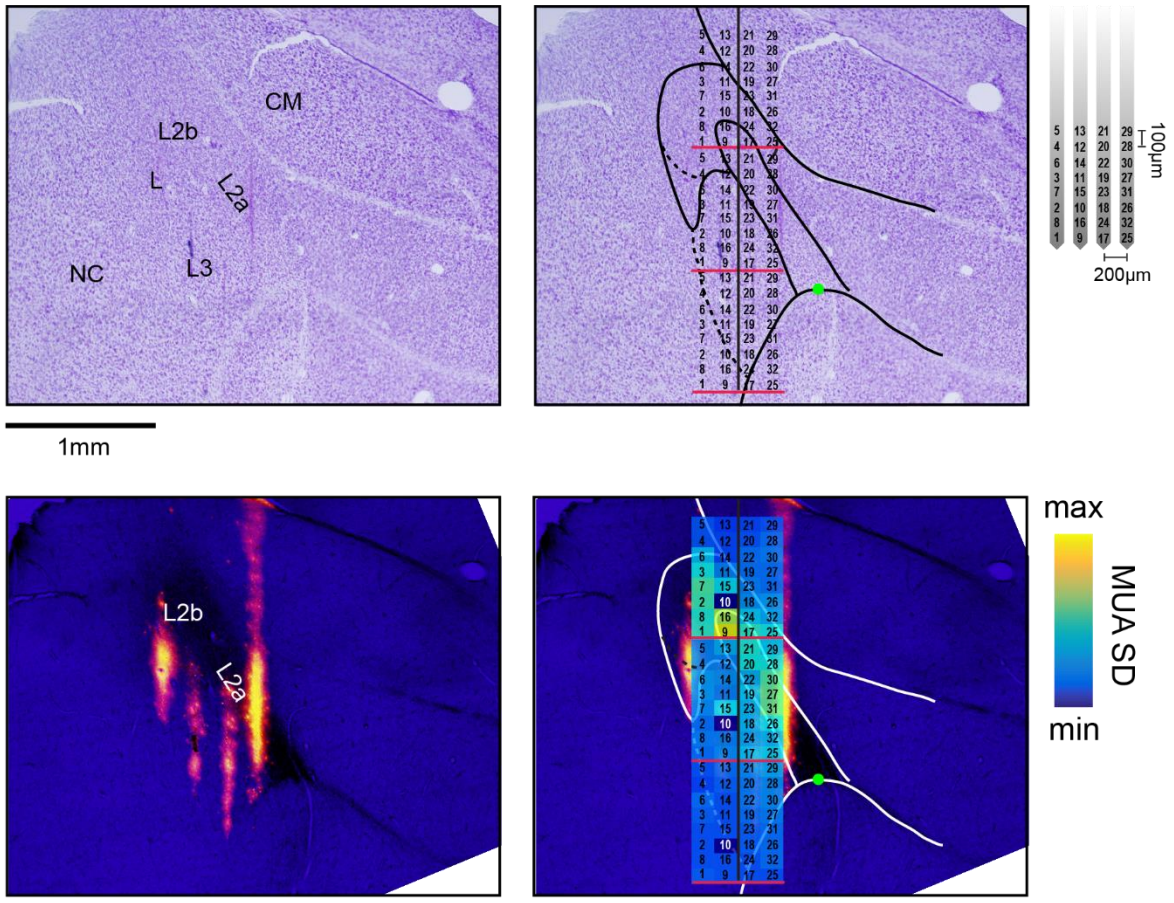

CM
L2b
L
L2a
NC
L3
1mm
100µm
200µm
L2b
L2a
max
MUA SD
min

Figure 2.4 Anatomical reconstruction of recording sites in the auditory cortex.
(A) Example bright-field microscopic images (4x) of parasagittal sections of the auditory cortex. Consecutive images of 40 μm sections are arranged from medial to lateral, ordered left to right and top to bottom. As sections progress from medial to lateral, the ventral tip of L2a increasingly intersects with the LMD and appears to fan out and fuse with the LMD. Green outline indicates the section that was selected as the medial-lateral reference plane, and green dots indicate the rostral-caudal and dorsal-ventral reference point. Reference points were determined for each hemisphere in each bird and used to standardize recording coordinates across birds. *(B)* Example Nissl (top) and bright-field images (bottom) taken at the reference plane. DiI signal appears as orange fluorescent tracks in the bottom two images. Microscopic images on the right are overlaid with traced lines delineating auditory regions, and the estimated position of numbered recording channels. Red lines indicate the depths at which recordings were conducted within the same penetration. The bottom-right image is overlaid with a heatmap showing the standard deviation of the multi-unit activity trace recorded from each channel (note that channel 10 was defective). L2a and L2b are characterized by relatively higher standard deviation than surrounding regions. The schematic on the top-right corner shows the geometrical arrangement of recording channels on the 32-channel multielectrode array. LMD, dorsal medullary lamina; MUA, multi-unit activity; SD, standard deviation.

2.5.5 Pre-processing of multichannel recording data

Data analysis was carried out in MATLAB (Mathworks). Spikes were detected and sorted offline using the WavClus automated sorting algorithm followed by manual refinement (Calabrese & Woolley, 2015; Quiroga et al., 2004). The WavClus algorithm is an unsupervised and fast method developed by researchers at the University of Leicester to detect and sorts spikes from multiunit recordings (Quiroga et al., 2004). To increase the signal-to-noise ratio (SNR), we applied a nonlinear filter on the bandpass-filtered voltage trace in order to emphasize high-amplitude and high-frequency voltage deflections. We then applied the WavClus algorithm to automatically detect and sort spikes on each channel, and manually refined the output by inspecting the waveform shape and amplitude of each cluster. Lastly, single units were identified on the basis of signal-to-noise ratio (SNR; the difference between mean of spike amplitudes and noise amplitudes divided by the geometric mean of their SDs), inter-spike interval distribution (the percentage of inter-spike-intervals shorter than 1 ms), and stability of recordings across trials. The 95% confidence interval of SNRs for single units in our dataset was 6.64 to 7.14, and that for the percentage of inter-spike intervals below 1 ms was 0.04% to 0.06%. This procedure identified a total of 1825 single units across L2a, L2b, L3 and NC.

Units were only included in the analysis of calls (vocoded and natural), ripple, and tone responses if they showed significant responses to at least 5% of all stimuli in each respective set. For the analysis of song responses, units were included if they showed significant responses to calls.

Significant responses were determined as following. Each unit's spontaneous firing rate was computed for the 200 ms periods preceding each trial. Driven firing rates were computed with spikes occurring between stimulus onset and 20 ms after stimulus offset. Onset firing rates were

computed with spikes occurring within the first 50 ms following stimulus onset. Evoked responses were considered significant if either the driven firing rate or the onset firing rate was significantly higher or lower than spontaneous rates at $p < 0.05$. This procedure yielded 847 call-responsive, 674 ripple-responsive, and 1184 tone-responsive units across L2a, L2b, L3 and NC.

2.5.6 Analyzing neural selectivity for spectral structure

Spectral selectivity index (SSI). SSI was computed from responses to vocoded calls with the following formula,

$$\mathrm{SSI} = \frac{\mathrm{FR}_{40,80} - \mathrm{FR}_{16,20}}{\mathrm{FR}_{40,80} + \mathrm{FR}_{16,20}}$$

where $\mathrm{FR}_{40,80}$ represents the average firing rate evoked by 40 and 80-channel vocoded calls, and $\mathrm{FR}_{16,20}$ represents the average firing rate evoked by 16 and 20-channel vocoded calls. Based on their SSI, units were segmented into low-resolution-selective (LS; SSI < -0.2), unselective (US; -0.2 < SSI < 0.2) and high-resolution-selective (HS; SSI > 0.2) groups. An SSI of 0.2 indicates 50% response enhancement to 40 and 80-channel calls compared to 16 and 20-channel vocoded calls, and an SSI of -0.2 indicates the opposite.

Response selectivity for call stimuli was defined as the proportion of natural and vocoded call stimuli that did not evoke driven firing rates significantly above spontaneous activity from a given unit. Response selectivity was computed for all call stimuli, as well as by stimulus type (for natural calls and vocoded calls by channel number).

2.5.7 Analyzing response dynamics to calls

Population peri-stimulus time histograms (PSTHs). Single-unit PSTHs were constructed by calculating the trial-averaged instantaneous firing rates in 1 ms bins and smoothing the responses with a 5 ms Hanning window. Population peri-stimulus time histograms (pPSTHs) were computed by averaging the raw or min-max normalized PSTHs across a population of single

units. For visualization purposes, pPSTHs were further smoothed by applying a 10 ms moving average.

First-spike latency. The latency of neural responses to a call stimulus was estimated by identifying the first time after stimulus onset at which spiking activity significantly deviated from spontaneous activity ($p < 0.05$), assuming that the neuron was firing spontaneously with Poisson statistics (Chase & Young, 2007; Schumacher et al., 2011). Since the onset ramps of natural and vocoded calls were not standardized, we defined each call stimulus' onset time as the time at which the sound pressure waveform amplitude first reached 5% of the maximum amplitude of the call, and quantified the time of first spike relative to this onset time. Call response latency was taken as the shortest latency among those computed for all natural and vocoded call stimuli.

We also related call response latencies to tone and ripple response latencies. Latencies of response to each tone and ripple stimulus was calculated using the same method as used to calculate latencies to call stimuli. Tone response latency of a neuron was determined by averaging latencies across sound intensities and taking the shortest average latency across tone frequencies. Ripple response latency of a neuron was taken as the shortest average latency across ripple depths, phases, and densities.

Onset index. Onset index was calculated from responses to natural and vocoded calls using the following formula,

$$\text{Onset Index} = \frac{\text{FR}_{\text{onset}} - \text{FR}_{\text{sustained}}}{\text{FR}_{\text{onset}} + \text{FR}_{\text{sustained}}}$$

where FR_{onset} represents the average firing rate during the first 50ms after stimulus onset, and $\text{FR}_{\text{sustained}}$ represents the average firing rate during the subsequent period until stimulus offset (Schumacher et al., 2011). For each neuron, onset indices were averaged across all call stimuli that

elicited a significant response (either the driven firing rate or the onset firing rate was significantly higher or lower than spontaneous rates at $p < 0.05$).

2.5.8 Analysis of pure tone responses

Pure tone responses were analyzed to obtain the best frequency (BF) and bandwidth (BW) of tone-responsive neurons. Response strength (evoked spike rate minus spontaneous spike rate) was computed for each frequency-level combination. A frequency response curve was computed by averaging response strengths across levels for each frequency. The BF of each neuron was taken as the frequency eliciting the maximum response. The BW of each neuron at each level was calculated by first obtaining the frequency response curve at that sound level, and then obtaining the width of the portion (s) of the curve that exceeds half of its peak value.

2.5.8 Analyzing responses to spectrally modulated ripples

Construction of depth-phase matrices. For each spectral modulation density (1.2, 1.6 and 2.0 cyc/kHz), a single unit's driven firing rates were computed for each depth-phase combination and used to construct a response matrix with depth on the Y axis and phase on the X axis (**Figure 2.5**). Averaging the depth-phase matrices within LS, US, and HS neurons resulted in group depth-phase matrices.

To construct population ripple response matrices, each units' depth-phase matrices were z-scored and averaged across the three spectral modulation frequencies. Standardized average depth-phase matrices for all units were averaged to obtain the population depth-phase matrix.

Best spectral modulation density. Best spectral modulation density was determined for each neuron by averaging the firing rates evoked by all ripples, across phases and depths, for each density, and selecting the modulation density with the highest average firing rate.

Best phase. A phase-response curve was generated by taking the average of responses across modulation depths at the neurons' best modulation density. The phase evoking the maximal response was taken as the neuron's best phase.

Modulation depth dependency. Modulation depth dependency of single-unit responses was computed via Spearman's tests of correlation between driven firing rates and ripple modulation depths at each tested modulation density and at the neurons' preferred phase (the phase eliciting maximal average response). A ρ of 1 indicates that firing rate increases monotonically with modulation depth, and a ρ of -1 indicates that firing rate decreases monotonically with increases in modulation depth.

Quantifying harmonicity of spectrally modulated ripples. The spectral theory of pitch extraction posits that the auditory system extracts the fundamental frequency of a sound by analyzing the neural pattern of activation along the tonotopic axis (Duifhuis et al., 1982; Scheffers, 1983; Shamma & Klein, 2000); the sound spectrum is compared to internally stored spectral templates, which contain the frequency component placement patterns for a range of possible fundamental frequencies. My harmonic template matching quantification of harmonicity was based on this idea. This analysis was applied to each ripple stimulus to estimate how well the stimulus power spectrum matches the spectral profile of harmonic sounds, where component frequencies are integer multiples of a common fundamental frequencies.

1) *Construction of harmonic templates*. A series of "harmonic sieves" were constructed with F0s varying from 300 to 1000 Hz in 5 Hz increments. For each sieve, "windows" were centered at the first 8 integer multiples of the F0 (F0*1, F0*2, F0*3 … F0*8), and shaped according to a Gaussian function with standard deviation of 30 Hz. The sieve

had a maximum value of one in each window and a value of zero outside of the windows.

2) *Obtain match quality with each harmonic template*. Each ripple was "passed through" all possible harmonic sieves by multiplying its frequency power spectrum (FPS) with the value of the harmonic sieve at each frequency. The multiplication products were then summed across frequencies to generate the match index, which represents how well the spectral profile of a ripple matched the harmonic template with a given fundamental frequency.

3) *Find best-matched template and determine match quality***.** The maximal match index across all F0s was taken as the harmonic template match index for a given ripple. The more stimulus frequency components passing through the best-matching harmonic sieve, and the more energy contained in each component passing through, the greater the resulting harmonic template match index.

Determining the spectral feature driving population responses to ripples. To investigate whether modulation depth and/or harmonicity explained the variance in population responses to ripple stimuli, we carried out comparisons of nested regression models.

To test the relationship between harmonic template match and population firing rates, we first conducted multiple linear regression with only modulation density and modulation depth as predictor variables. Then, we conducted a second multiple linear regression, with harmonicity as an additional variable. An F test was used to compare the second model (with 3 predictor variables, plus an intercept) to the first one (with 2 predictor variables, plus an intercept), which indicated whether the addition of harmonic template match significantly increased the predictive power of the model. To test the relationship between modulation depth and population firing rates, we

carried out the same procedure, except that the first model included modulation density and harmonicity as predictor variables, and modulation depth was added in the second model.

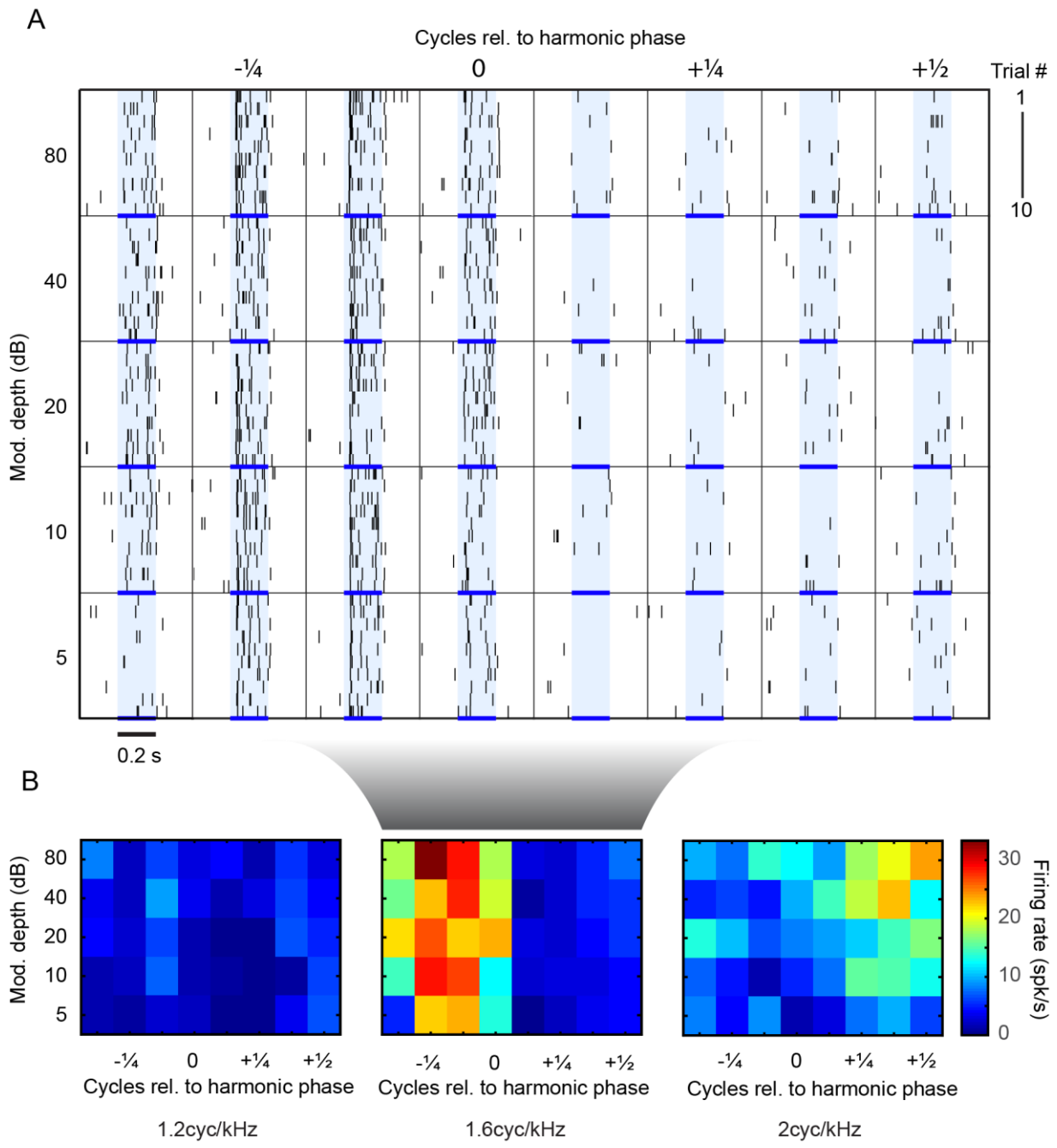


Figure 2.5 Construction of depth-phase ripple response matrices.
(A) Raster plots showing an example single neuron's spikes aligned to presentations of ripple stimuli with different modulation depths, varied along the Y axis, and phases, varying along the X axis. Blue bars and shading indicate duration of stimulus presentation. Data shown are responses to ripples with modulation density of 1.6 cyc/kHz *(B)* Depth-phase matrices in which each pixel is colored-coded according to the firing rate evoked by a ripple of the corresponding depth-phase combination. The best modulation density of this neuron is 1.6 cyc/kHz.

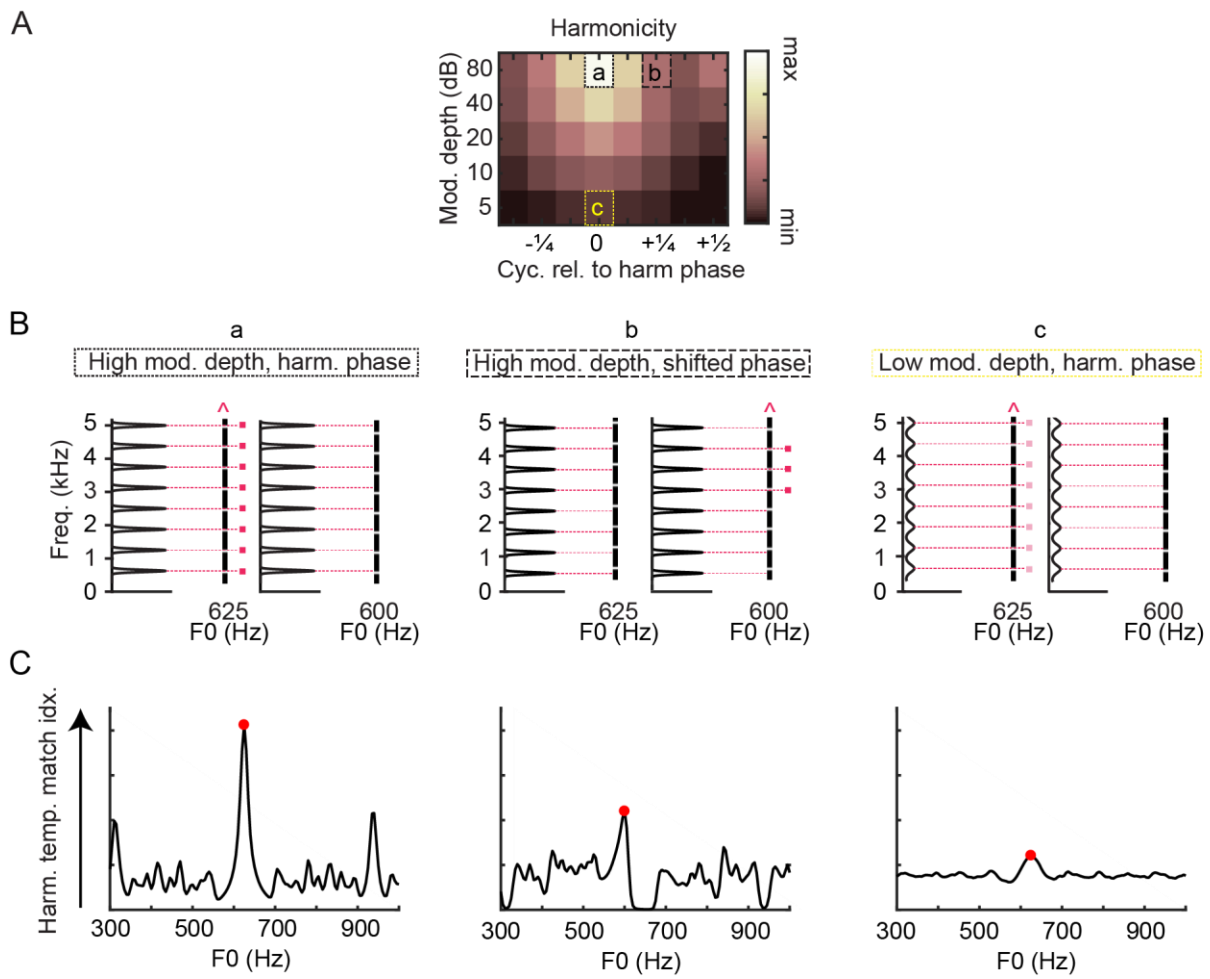


Figure 2.6 Quantification of harmonicity by harmonic template matching
(A) Heat map showing how harmonicity of ripple stimuli varies with spectral modulation depth and phase. Squares outlined by black and yellow dashed lines indicate correspondence between shown ripples' position in the depth-phase matrix and its harmonic template match schematic in *(B)*. *(B)* Schematic of harmonic template analysis. Spectral profiles (left) of three representative example ripples with modulation density of 1.6 cyc/kHz and their alignment with two example harmonic templates/sieves (right) are shown. Note that only frequencies up to 5 kHz are shown. Rose-colored ^ symbols indicate the sieve for which the ripple was the best match. For the most closely matched sieve, rose-colored squares indicate positions where spectral peaks of the stimulus "falls through" the sieve. *(C)* Relative harmonic template match index for harmonic sieves with different fundamental frequencies. Red filled circles indicate the maximum match index and the corresponding best-match harmonic sieve. The high modulation depth and harmonic phase ripple (left) matches the 625 Hz template best and has the maximal match index. The high modulation depth and shifted phase ripple (middle) matches the 600 Hz template best and has intermediate match index. The low modulation depth and harmonic phase ripple (right) matches the 625 Hz template best and has the lowest match index.

2.5.9 Analyzing population responses to songs

Population peri-stimulus time histograms (pPSTHs). Single-unit song PSTHs evoked were obtained by calculating the trial-averaged instantaneous firing rates in 1 ms bins. pPSTHs were computed by averaging single unit PSTHs across a population of single units. pPSTHs were smoothed by applying a 15 ms moving average. To examine response transformations between auditory regions, we computed pPSTHs to each song for the L2a, L2b, L3, and NC populations. To examine how spectral selectivity for vocoded calls translates to the encoding of song, we computed pPSTHs separately for LS, US, and HS neurons in L3.

We quantified how fast each regions' and each cell types' population responses increased and decreased with syllable onsets and offsets within songs. Syllable onsets and offsets of each song were determined by applying an amplitude threshold on the smoothed and normalized absolute values of song waveforms. To determine population response rise time, pPSTHs were first normalized to range from 0 to 1. The pPSTH segment from each syllable onset until 20 ms after stimulus offset was obtained, and the earliest peak in population response was determined using the MATLAB function findpeaks( ) with minimum peak prominence of 0.1. The timing of this earliest pPSTH peak relative to syllable onset was taken as the rise time for each syllable. Population response fall times were determined similarly using the MATLAB findpeaks( ) on the negative image of the pPSTH segments during the 100 ms after syllable offsets. The timing of the earliest pPSTH trough relative to syllable offset was taken as the fall time for each syllable.

Analyzing time-varying song acoustics. In order to relate population responses to time-varying acoustic features, we calculated the goodness of pitch, frequency modulation, mean frequency, and Wiener entropy of each song using the Sound Analysis Tools (SAT) for MATLAB package (Tchernichovski et al., 2000). The SAT package is the MATLAB

implementation of the Sound Analysis Pro (SAP) software, which has been widely used as a standardized method to analyze the acoustic features of zebra finch song. The SAP and SAT methods makes acoustic measurements based on frequency derivatives, which serves as “edge detectors” of frequency traces in a spectrogram. While a traditional spectrogram shows the power of sound in each time-frequency combination, the frequency derivative shows the change of power. The four acoustic features that we measured are described below.

Goodness of pitch is defined as the peak of the derivative-cepstrum. It measures how periodic or harmonic a sound is. The cepstrum is the result of taking the inverse Fourier transform of the logarithm of the spectrum of a sound. The cepstrum has been used for the detection of voiced and unvoiced speech and for the detection of F0 in speech (Noll, 1964); unvoiced signals show no peak in its cepstrum, whereas voiced signals show a cepstral peak at the fundamental period. Noisy sounds (with relatively flat spectra) and pure tones give low values for goodness of pitch, while harmonic stacks give high values.

Frequency modulation is a measure of how fast frequencies change over time; it estimates the slope of the frequency traces with respect to a horizontal line. Mean frequency represents the center of distribution of power across frequencies, calculated by estimating the central tendency of the derivative power distributions. It provides a smooth estimate of where the power in the sound is concentrated.

Wiener entropy measures how wide and uniform the sound spectrum is. It is defined as the ratio of the geometric mean to the arithmetic mean of the sound spectrum. Entropy is expressed on a logarithmic scale, ranging from 0 (white noise) to negative infinity (complete order in the spectrum, such as a pure tone). Noisy sounds, with sound energy smeared within the frequency range, are more uniform in their spectrum and give values close to zero; harmonic

stacks, which contain organized peaks and valleys and thus are less uniform in their frequency structure, give moderately negative values; tonal sounds give larger negative values.

For each of the 4 acoustic features described above, a feature vector sampled at 1 ms intervals was extracted for each song. Feature vectors were smoothed with the same time window as for pPSTHs, and used to relate neural responses to acoustic features. In order to avoid including silent segments in acoustic analysis, SAT acoustic features were analyzed only for segments of song where the amplitude envelop exceeded 5% of the maximum amplitude of the song.

2.6 CONCLUSIONS

In this chapter, I described methods used to study how the auditory system functions during social communication. A combination of natural vocalizations, vocalization-like stimuli, and synthetic stimuli were curated for use in behavioral and neural experiments. Natural vocalizations have known behavioral value and function to the animal, which provided us the ability to establish a behavioral baseline for assessing the effectiveness of vocalization-like stimuli in driving natural behaviors. Vocalization-like stimuli were generated using natural vocalizations as templates, with specific spectral manipulations to identify critical acoustic cues for social communication. Using a call-and-response behavioral paradigm, we identified behaviorally relevant spectral parameters of vocalizations, which motivated us to study how these parameters were processed by the auditory system. In our neurophysiological experiments, we utilized multichannel recording methods that enabled us to map the spatial organization of response properties in the brain. Emphasis was placed on identifying consistent landmarks permitting the mapping of recording sites and standardizing the measurement of anatomical locations across animals. Our neural response analyses focused on separating two commonly confounded spectral parameters of vocalizations using synthetic stimuli,

then using synthetic stimuli responses to predict neurons' encoding of natural vocalizations. Chapters 3 and 4 will describe our identification of spectral parameters of vocalizations important for social communication, and our findings on how the auditory cortex processes these behaviorally relevant acoustic features.

Chapter 3

BEHAVIORAL AND NEURAL SELECTIVITY FOR THE SPECTRAL STRUCTURE OF VOCAL SOUNDS

3.1 ABSTRACT

Vocal communication relies on the ability of listeners to identify, process, and respond appropriately to vocal sounds produced by others in complex environments. In order to accurately recognize these social signals, animals' auditory systems must robustly represent acoustic features that distinguish vocal sounds from other sounds in the environment. The spectra of vocalizations contain certain structural features that could contribute to auditory processing and extraction of vocal signals. Vocalizations typically contain spectral modulations, or regular fluctuations in power along the frequency axis. Spectral modulation is closely related to harmonicity, which refers to spectral energy concentrated at integer multiples of a fundamental frequency. Harmonic sounds give rise to the perception of pitch and constitute vocalizations of animals ranging from humans to frogs. Reduction of spectral information into a limited number of channels, a manipulation that diminishes both spectral modulation and harmonicity, affects human listeners' perception of social information from these signals. The neural mechanisms underlying the perceptual reliance on spectral resolution are not well understood.

Here, we test the role of vocalization-typical spectral features in behavioral recognition and neural processing of vocal sounds, using songbirds. We found that the spectral resolution of natural communication calls must be preserved to a certain degree to elicit vocal responses from songbirds. To elicit responses, call stimuli must contain distinct spectral peaks and valleys, but the spectral

peaks need not be harmonically related. We further identify a population of neurons in the deep region of auditory cortex that represents a neural correlate of birds' behavioral sensitivity to the spectral resolution of calls.

3.2 INTRODUCTION

Vocal communication relies on auditory processing of vocalizations, which convey social information to listeners (Belin et al., 2004; Seyfarth & Cheney, 2017). Vocalizations are composed of characteristic acoustic features that distinguish them from other sounds in the environment (Attias & Schreiner, 1997; Rieke et al., 1995; Singh & Theunissen, 2003; S. M. N. Woolley et al., 2005). To enable effective processing of vocalizations for social communication, the auditory system may be tuned to the acoustic signatures of vocal sounds.

The spectral features that distinguish vocalizations from other sounds may facilitate their processing and perception. Vocal sounds generally contain regular fluctuations in power across the frequency axis, a property known as spectral modulation (Singh & Theunissen, 2003). Regular spectral modulation along the linear frequency axis results in distinct and evenly spaced peaks and valleys in the sound spectrum. In natural sounds, spectral modulations are closely tied to harmonicity, which refers to simultaneous frequency components at integer multiples of a fundamental frequency (F0) (X. Wang, 2013; X. Wang & Walker, 2012). Harmonic vocal sounds result from periodic oscillations of the vocal folds, whose frequency of vibrations determine the F0 (Riede & Goller, 2010; Titze, 2017). Such sounds constitute human speech and the communication vocalizations of other animals, including frogs, birds, bats, cats, elephants, and non-human primates (Simmons & Simmons, 2011; Soltis, 2010; X. Wang, 2013). Harmonicity gives rise to the perception of pitch, and is thought to contribute to auditory object identification (Walker et al., 2011).

In order for a signal to convey the typical spectral structure of vocalizations, including spectral modulations and harmonicity, a certain level of spectral resolution must be achieved. In other words, the frequency axis must be sampled at sufficiently small bins, in order to reconstruct the spectral details that characterize vocal sounds. The perceptual relevance of spectral structure has been studied in the context of human speech perception and is reviewed in Section 1.3.3.

In animal models, neural and behavioral discrimination of complex sounds appear to be robust to spectral degradation. In Mongolian gerbils (*Meriones unguiculatus*), primary auditory cortex neurons were able to distinguish between different categories of communication calls even with substantial spectral degradation (frequencies from 2 to 40 kHz reduced to 4, 8, or 16 channels) (Ter-mikaelian et al., 2018). In a study of European Starlings (*Sturnus vulgaris;* further reviewed in Section 1.4.3), birds were able to recognize sound sequences despite spectral degradation (Bregman et al., 2016). While previous studies looked at whether spectral details were used for sound classification and recognition, our studies addressed whether they contribute to the behavioral relevance of vocalizations. We defined behaviorally relevant sounds as those that evoke vocal responses from the listener. By determining the behavioral relevance of spectrally altered calls, we aimed to identify the spectral cues animals use in vocal communication.

In this chapter, I detail behavioral experiments examining the role of two different aspects of spectral structure in vocal communication of the zebra finch (*Taeniopygia guttata*; further reviewed in Section 1.6), a social songbird used as a model system for studies of vocal communication and learning (Brainard & Doupe, 2013). I also describe neurophysiological studies where we identified a neural correlate for behavioral sensitivity to the spectral resolution of vocalizations.

3.3 RESULTS

3.3.1 Birds produce distinct call types in call-and-response experiments

Zebra finches exchange distance calls when visually separated (Zann, 1996), and birds recognize their mates' voices using these calls (Vignal et al., 2004). Socially-isolated birds readily respond to distance call playbacks by vocalizing (Perez et al., 2015; Vicario et al., 2001; Vignal & Mathevon, 2011). The "call and response" behavioral test assessed how birds' responses differed with variations in the spectral properties of acoustic stimuli.

The vocalizations that birds produced during call-and-response experiments were not homogenous. We identified distinct types of vocalizations from audio recordings, which contained all sounds produced by birds within the first 10 s after stimulus onset. We extracted all call vocalizations by identifying their onset and offset times within a trial, and then used these timestamps to extract calls for acoustic analysis.

When vocalizations of each bird were plotted along the dimensions of mean frequency and duration, two or more clusters were present (**Figure 3.1A** and **Figure 3.2A**). Previous studies have shown that distance calls have higher mean frequencies and are longer in duration than other types of contact calls (Elie & Theunissen, 2016). To isolate distance calls from other call types, we performed k-means clustering for each birds' vocalizations based on mean frequency (measured using SAT, see Section 2.5.9) and duration (measured by subtracting response call onset time from offset time). The number of clusters for each bird, ranging from 2 to 3, was chosen by visual inspection of the joint mean frequency-duration distribution of the birds' vocalizations. After vocalizations were assigned to clusters, the cluster with greater average mean frequency and longer mean duration was taken to represent distance calls.

Vocalizations that were not classified as distance calls could correspond to *Tet* or *Stack* calls, which are acoustically distinct categories (Elie & Theunissen, 2016; Maat et al., 2014). However, we treated all non-distance calls as a single group of "short calls" for two reasons. First, the birds in our experiments did not consistently produce two distinct shorter call types. Second, a previous study has shown that these two call types were performed in the same behavioral contexts: in a seemingly automatic and continuous fashion, when zebra finches move around in their surroundings (Elie & Theunissen, 2016).

In order to verify that distance call responses were the appropriate vocalization category to treat as stimulus-evoked responses, we compared distance calls and short calls in two aspects: their time course relative to stimulus presentation, and whether birds' production of these calls after playbacks of noise differed from that after playbacks of other birds' calls.

We found that distance calls and short calls had different time courses. The average number of distance calls showed a strong peak within the first second after presentation of stimuli, while the number of short calls were relatively evenly distributed across time after stimulus presentation (**Figure 3.1B** and **Figure 3.2B**). This shows that birds' production of distance calls was likely a direct response to preceding stimulus, while short calls were emitted continuously regardless of stimulus presentation.

In addition to differences in their timing, distance calls and short calls also differ in stimulus dependency. Birds produced distance calls in greater numbers (**Figure 3.1C** and **Figure 3.2C;** paired t-tests, $p < 0.05$) and in a greater proportion of trials (**Figure 3.1D** and **Figure 3.2D;** paired t-tests, $p < 0.05$) to natural calls than they did to noise stimuli. In contrast, the number of short calls produced and proportion of trials with short calls did not differ between noise and natural call

trials. This shows that distance calls constituted a specific response to social stimuli, while short calls were not modulated by stimulus type.

In the two experiments that we conducted with largely non-overlapping birds (**Figure 3.1-3.2)**, we identified the same distinct call types using the same acoustic features. In addition, distance calls and short calls differed in time course and stimulus dependency in the same ways across these two experiments. Thus, these behavioral phenomena are robust and replicate across studies.

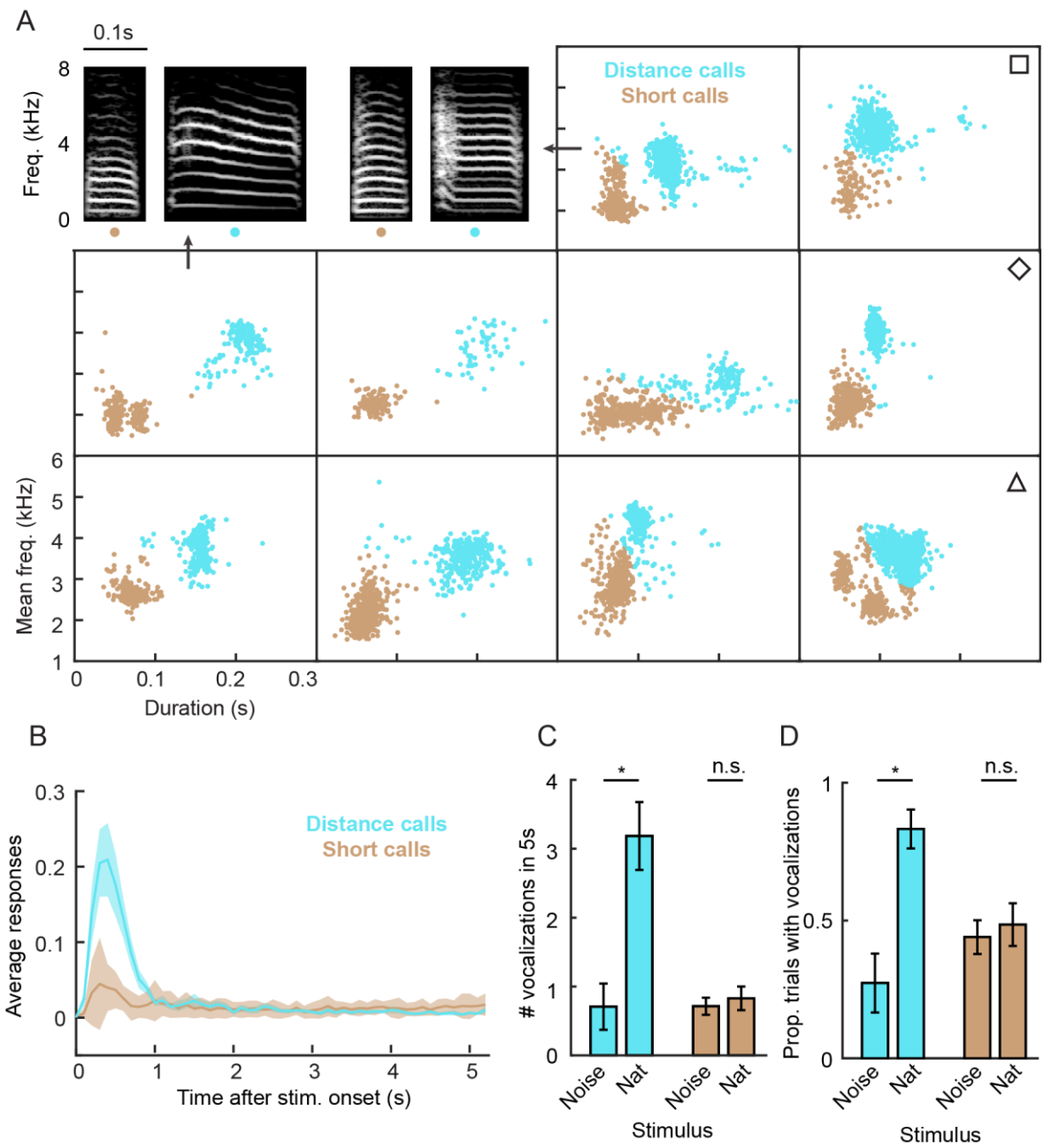


Figure 3.1 Time course and stimulus dependency of distance call responses.
(A) Scatter plots showing the mean frequency and duration of all recorded calls during Behavioral Experiment 1. We acoustically analyzed vocalizations within the first 10s after stimulus onset. Each plot shows data from one bird, and each data point represents one vocalization. Square, diamond, and triangle outlines on the top-right corner of scatter plots denote the birds that were used in both behavioral experiments and correspond with birds shown in **Figure 3.2.** Vocalizations formed clusters and distance calls (blue) were identified as the cluster with higher mean frequency and duration. Other vocalizations had lower mean frequencies and shorter durations, and were collectively classified as "short calls" (tan). Spectrograms show representative examples of distance calls (blue) and short calls (tan) respectively. *(B)* Average number of distance calls (blue) and short calls (tan) in the first 5s

following the onset of stimulus presentation, presented in 100ms time bins. The production of distance calls showed a prominent peak within the first second, while the production of short calls was less temporally locked to stimulus onset. *(C, D)* Average number of vocalizations *(C)* and proportion trials with at least one vocalization *(D)* in response to noise or natural call stimuli, shown separately for distance calls and short calls (mean ± SEM, N=10 birds). Birds produced more distance calls and produced them in a greater proportion of trials in response to natural calls than to noise. In contrast, the number of and proportion of trials with short calls did not differ in response to noise and natural calls (* $p < 0.05$, paired t test).

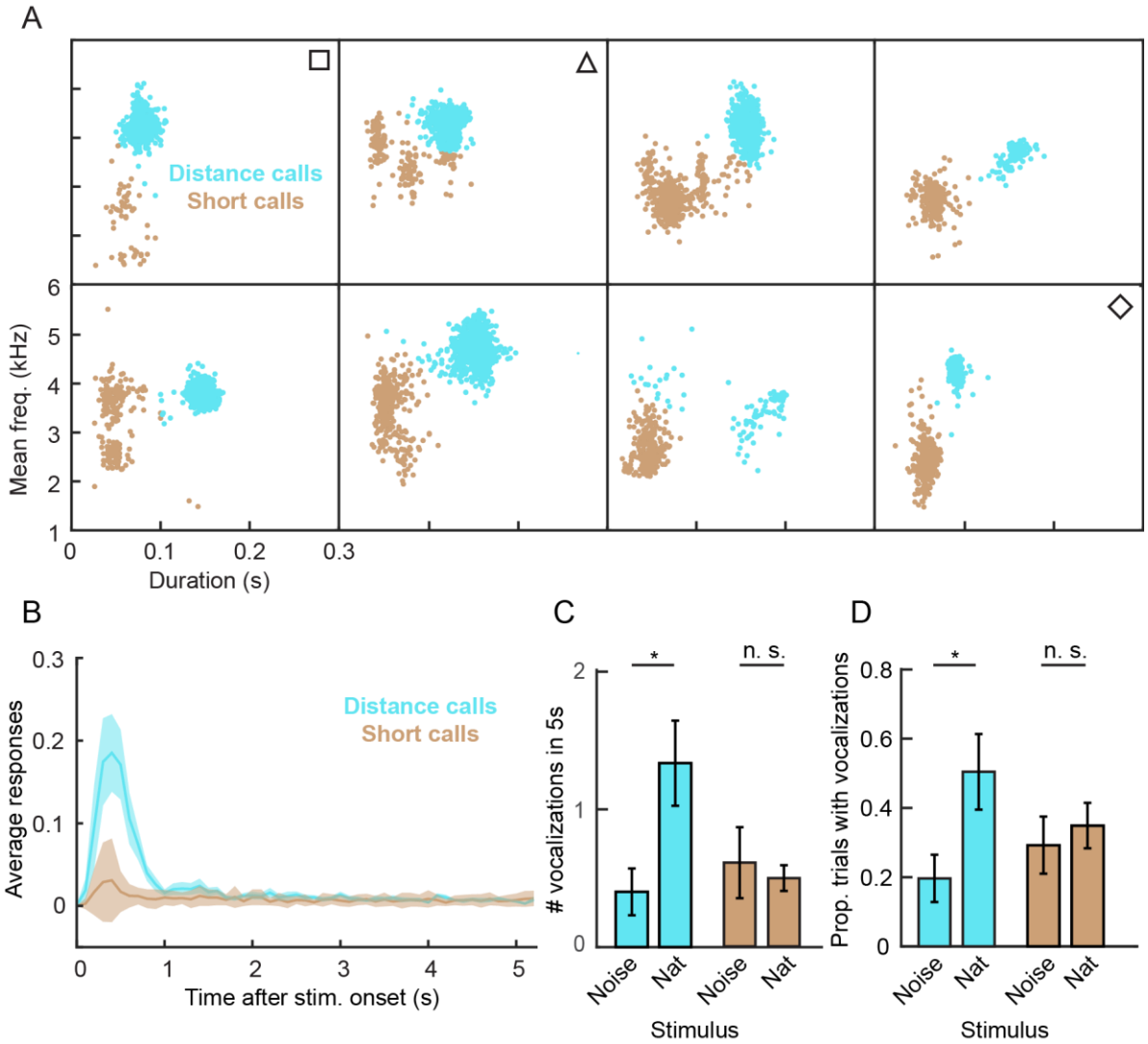


Figure 3.2 Replication of time course and stimulus dependency of distance call responses. *(A)* Scatter plots showing the mean frequency and duration of all recorded calls during Behavioral Experiment 2. Data presented as in **Figure 3.1A**. Square, diamond, and triangle outlines on the top-right corner of scatter plots denote the birds that were used in both behavioral experiments and correspond with birds in **Figure 3.1.** *(B)* Average number of distance calls (blue) and short calls (tan) in the first 5s following the onset of stimulus presentation, presented in 100ms time bins. The production of distance calls showed a prominent peak within the first second, while the production of short calls was less temporally locked to stimulus onset. *(C, D)* Average number of vocalizations *(C)* and proportion trials with at least one vocalization *(D)* in response to noise or natural call stimuli, shown separately for distance calls and short calls (mean ± SEM, N = 8 birds). Birds produced more distance calls and produced them in a greater proportion of trials in response to natural calls than to noise. In contrast, the number of and proportion of trials with short calls did not differ in response to noise and natural calls (* $p < 0.05$, paired t test).

3.3.2 Behavioral experiment 1: the effect of spectral degradation on vocal responses to communication calls

We first tested how spectral degradation affects the behavioral relevance of vocal communication sounds. Birds reliably responded to presentations of natural calls by producing their own distance calls (**Figure 3.3A**). The occurrence of response calls peaked within 1 s of stimulus call onset. The average strength of responses to natural calls was significantly higher than the average strength of responses to noise in the first 3 s (**Figure 3.3B,** two sample t-tests, $p < 0.01$). Responses to noise and natural calls provided the behavioral baseline against which the effects of spectral manipulations on response calls could be assessed.

To determine whether spectral degradation affects birds' vocal responses, we created vocoded calls that varied in spectral resolution (noise-vocoding procedure detailed in Section 2.2.2). Briefly, vocoded calls were composed of linearly spaced spectral channels. Each channel consisted of band-pass filtered noise whose amplitude modulation matched that of the corresponding natural call. The presence of distinct and evenly spaced frequency components was apparent in vocoded calls with 40 and 80 channels, as spectral channels were narrow enough for adjacent frequency components to fall into different channels (**Figure 3.3C**).

We presented natural calls, vocoded calls, and filtered noise segments to isolated male birds. With increasing spectral resolution (channel number), the probability of a bird responding with a distance call (**Figure 3.3D**) and the number of response distance calls (**Figure 3.3E**) increased. The effect of stimulus type on both behavioral measures was significant (repeated-measures ANOVA, proportion trials: $F(6, 54) = 26.92$, $p < 0.001$; no. of response calls: $F(6, 54) = 25.73$, $p < 0.001$). Birds responded to a greater proportion of vocoded calls with 40 and 80 spectral channels than to noise (**Figure 3.3D**, $p < 0.001$, Tukey's test). The 80-channel vocoded calls

elicited more response calls than did noise (**Figure 3.3E,** $p < 0.001$, Tukey's test). For both behavioral measures, natural calls elicited significantly more responses than did vocoded calls or noise (all $p < 0.05$, Tukey's test). These results showed that birds responded more to vocoded calls with higher spectral resolution, indicating that the presence of discrete frequency components was necessary for eliciting behavioral responses.

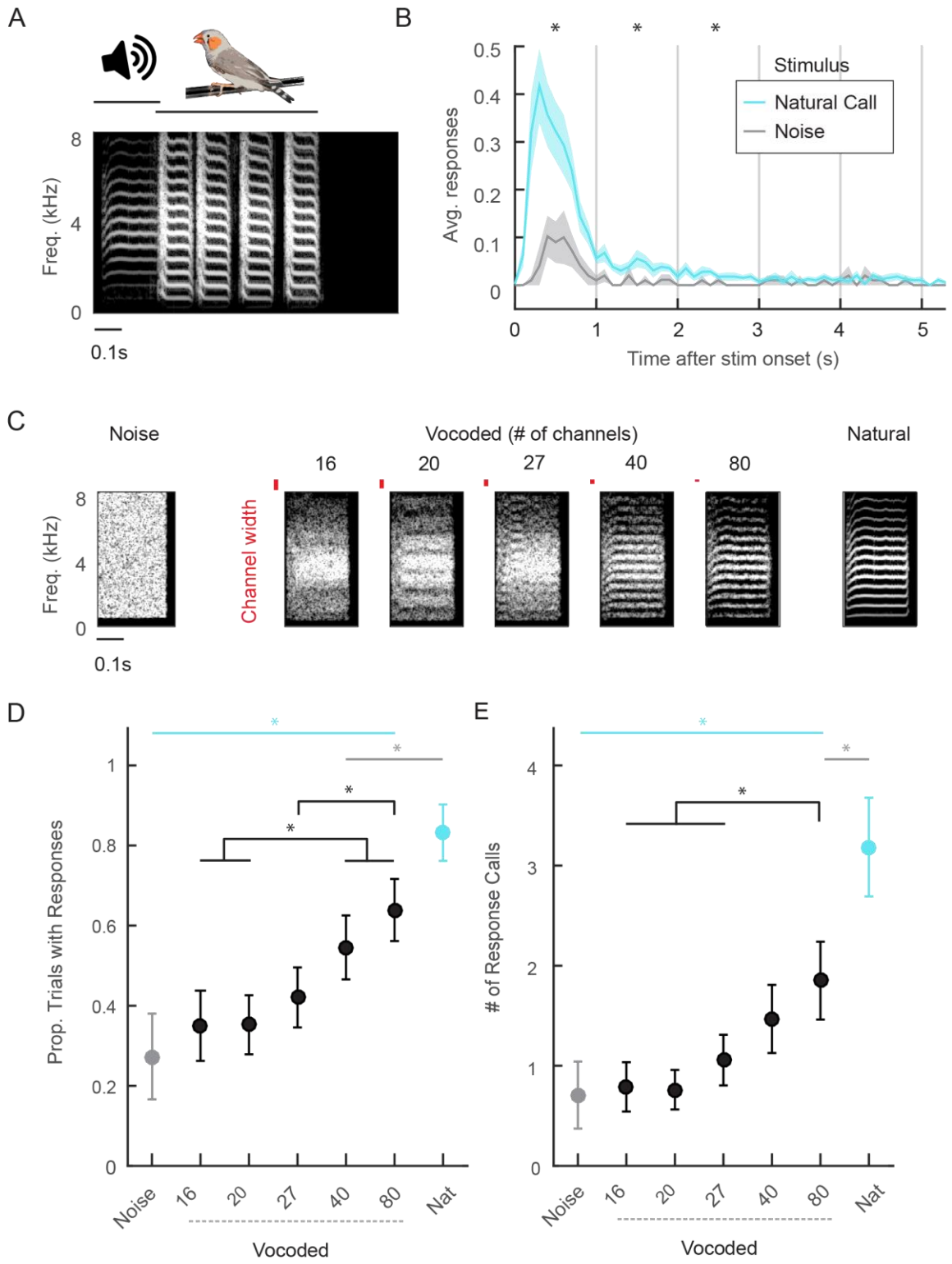


Figure 3.3 Spectral degradation decreases vocal responses to calls.

(A) Example of a single trial in the call-response behavior paradigm, in which presentation of a natural call evoked four response calls. *(B)* Average number of response calls in the first 5s following the onset of natural call and noise presentation, shown in 100ms time bins. Asterisks indicate significant differences between responses to natural calls and noise for each 1s period (* $p < 0.05$, two sample t-tests). *(C)* Example spectrograms of noise, vocoded versions of a call and the natural call show the differences in acoustic structure across test stimuli. Red bars indicate the width of each channel for each vocoded call shown. *(D)* Proportion of trials in which birds produced at least one response call (left) and *(E)* average number of response calls in the first 5 s for each stimulus type (mean ± SEM, N=10). Cyan and gray asterisks indicate significant differences from natural calls and noise respectively in stimulus-evoked responses. Black asterisks and brackets denote significant differences between responses evoked by vocoded calls with different channel numbers (* $p < 0.05$, Repeated-measures ANOVA with Tukey tests).

3.3.3 Behavioral experiment 2: the role of harmonicity in vocal responses to communication calls

Spectral degradation with 27 or fewer channels reduces both modulation depth (distance between peaks and valleys in the call spectrum) and harmonicity (presence of frequency components that are integer multiples of a F0). To determine whether harmonicity was a salient feature for eliciting behavioral responses, we tested birds with inharmonic calls in which harmonic frequency ratios between spectral peaks were disrupted, but the discrete spectral components were maintained (**Figure 3.4A**). For inharmonic calls, each frequency component was randomly shifted up or down by a maximum amount of 10%, 30%, or 50% of the F0.

Using the "call and response" paradigm, we presented natural calls (Nat), synthesized harmonic calls (Harm), synthesized inharmonic calls (with maximum frequency shifts of 10%, 30% or 50%), and filtered noise segments. As in the previous behavioral experiment, stimulus type had a significant effect on birds' responses (repeated-measures ANOVA, proportion trials: $F(5, 35) = 12.97$, $p < 0.001$; no. of response calls: $F(5,35) = 9.8$, $p < 0.001$). Responses to inharmonic calls, regardless of maximum frequency shift, did not differ significantly from responses to Harm and Nat calls (**Fig 3.4B-C**; Tukey's tests, all $p > 0.9$ for proportion trials and no. of response calls). Responses to all inharmonic calls were significantly above noise-evoked responses (**Fig 2B-C**, Tukey's tests, all $p < 0.001$). These results show that birds' responses to communication calls did not depend on harmonicity.

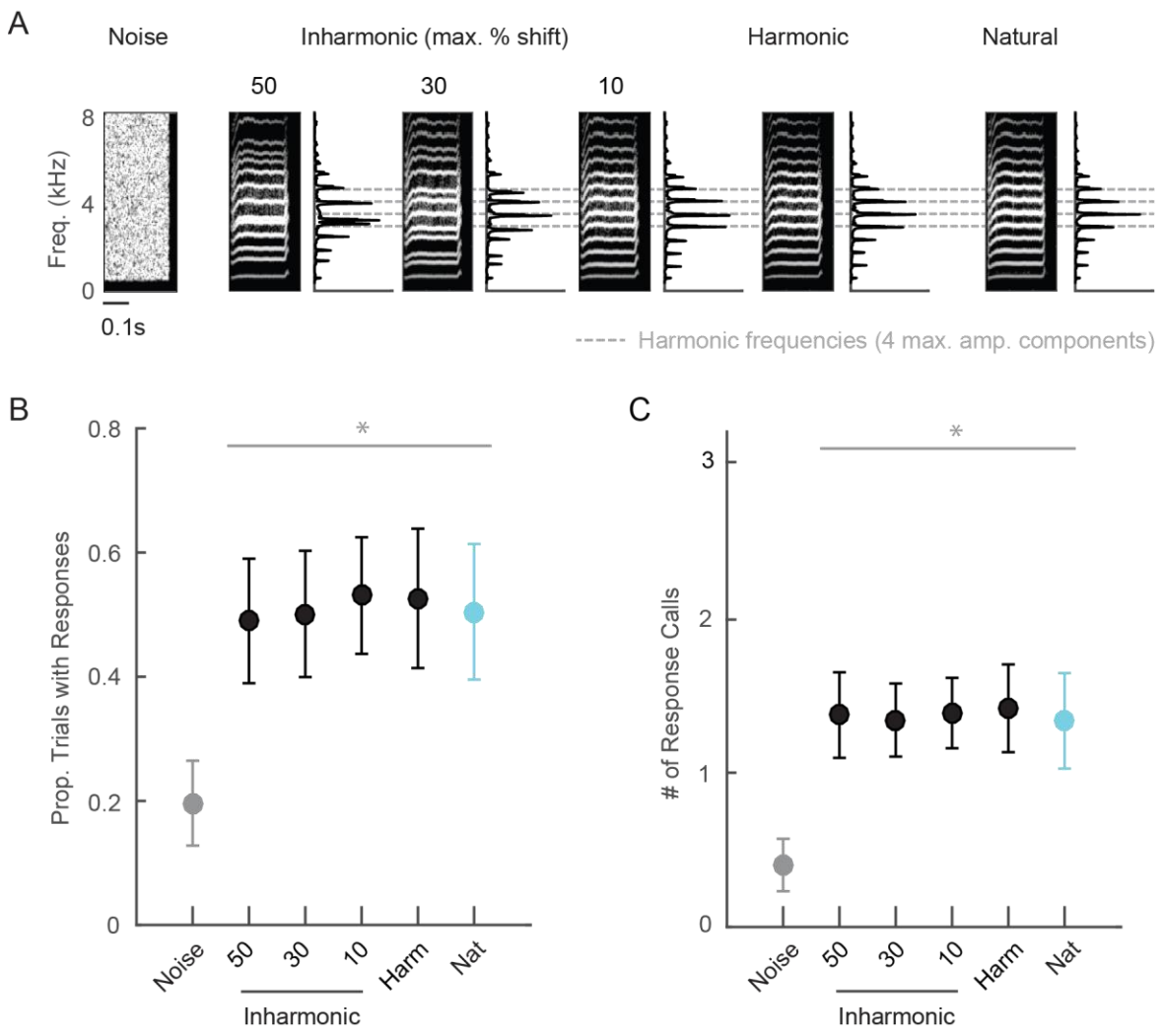


Figure 3.4 Vocal responses to calls are not affected by inharmonicity.
(A) Example spectrograms of noise, synthesized inharmonic calls, synthesized harmonic call, and natural call. Gray dashed lines indicate harmonic placement of frequencies for the 4 components with highest amplitude. *(B)* Proportion of trials in which birds produced at least one response call and *(C)* average number of response calls for each stimulus type (mean ± SEM, N=8). Gray asterisks indicate significant differences from noise in stimulus-evoked responses (* $p < 0.05$, Repeated-measures ANOVA with Tukey's tests).

3.3.4 Call responses show hierarchical progression in the auditory cortex

Because we found that birds' responses to communication sounds were dependent on spectral resolution, we sought to identify neurons in the auditory cortex whose responses were similarly modulated by spectral resolution. Using 32 channel multi-electrode arrays, we recorded the extracellular activity of single auditory neurons in awake birds presented with natural and vocoded calls (**Figure 3.5B).** Similar to mammalian cortex, zebra finch auditory cortex (AC) processes sound hierarchically (**Figure 3.5A-B).** Field L intermediate regions, L2a and L2b (hereafter intermediate-a and intermediate-b), receive input from the auditory thalamus and relay information to the superficial (L1/CM) and deep (L3) regions; the superficial regions also project to the deep region (Vates et al., 1996). The deep region is a major source of projections to the secondary auditory cortex (NC) and subcortical regions (Mello et al., 1998).

Single neurons' spiking responses to calls (natural and vocoded) were progressively lower and more selective along the cortical processing pathway (**Figure 3.5C**), in agreement with previous reports of song encoding in these regions (Calabrese & Woolley, 2015; Meliza & Margoliash, 2012). Stimulus-evoked firing rates differed significantly across brain regions (**Figure 3.5C, Left,** ANOVA, $F(3,843) = 23.09$, $p < 0.001$); neurons in the intermediate-a region fired at higher rates than did neurons in all other regions (Tukey's test, $p < 0.001$), and secondary-region neurons fired at lower rates than did neurons in other regions (Tukey's test, $p < 0.05$). Response selectivity, defined as the proportion of calls (natural and vocoded) that failed to evoke significant driven firing rates, was significantly lower in the intermediate-a region than in all other regions (**Figure 3.5C, Right,** ANOVA, $F(3, 843) = 13.11$, $p < 0.001$; Tukey's test, $p < 0.001$).

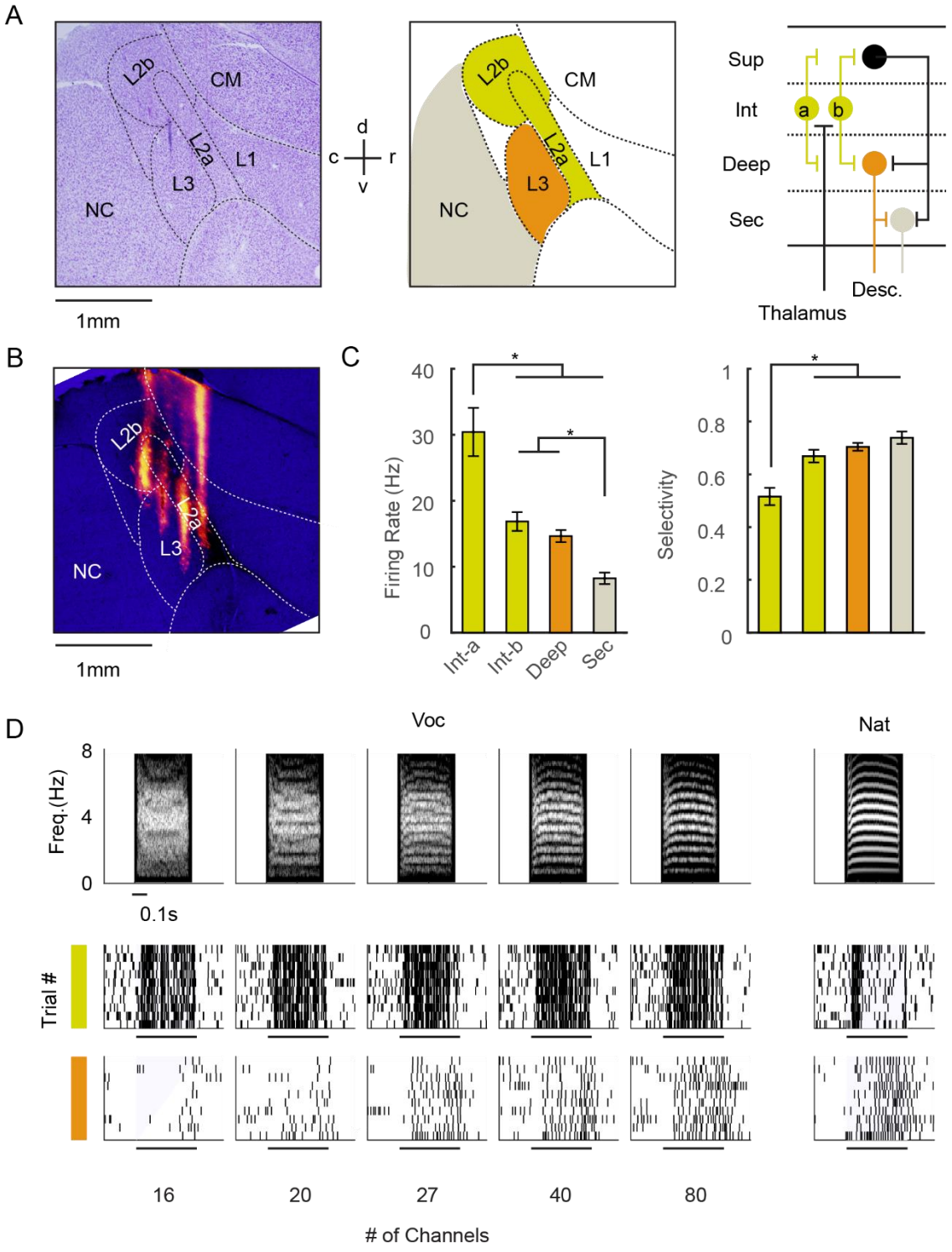


Figure 3.5 Anatomical organization, major connections, and neural responses of the songbird auditory cortex (AC).
(A, Left) Cresyl violet stained parasagittal section of the songbird brain. Dashed lines delineate major anatomical subdivisions. *(Middle)* Traced diagram of the same section, with color and

labeling denoting cortical regions. *(Right)* Circuit diagram of major projections in the songbird AC. Adapted from: (Calabrese & Woolley, 2015; Y. Wang et al., 2010) *(B)* Histological section showing locations of the four electrode shanks, labeled with DiI (orange). Electrode shanks show recording sites spanning multiple AC regions. *(C)* Average driven firing rates *(left)* and selectivity *(right)* in different regions computed from single neuron responses to natural and vocoded calls (* $p < 0.05$, one-way ANOVA with Tukey tests; intermediate-a, N = 111; intermediate-b, N = 189; deep, N = 411; secondary, N = 136 call-responsive units). Bar graphs show mean ± SEM. *(D)* Example spike trains of single neurons in intermediate and deep regions illustrating responses evoked by vocoded and natural calls. Spectrograms of stimuli are shown above the spike trains.

3.3.5 Auditory cortex tonotopy differs along the medial-lateral axis

Tonotopy, the systematic variation in best frequencies (BFs) of neurons from low to high values along a spatial axis, is a fundamental organizing principle of the primary auditory cortex (Bendor & Wang, 2005; Terleph et al., 2006; Tsukano et al., 2017). We sampled units broadly in the auditory cortex, using electrode arrays where 32 recording sites were arranged in a dense grid with known positional differences from each other. This allowed us to investigate the extent of tonotopic organization and relate it to spectral resolution selectivity. We anatomically mapped BFs along the sagittal plane by plotting neurons by their reconstructed recording coordinates, color coding them by best frequency, and overlaying these plots on anatomically traced sagittal sections of the auditory cortex (**Figure 3.6A**).

In the medial portion of auditory cortex (estimated 0.7 mm – 1.3 mm from the midline), neurons were organized by BFs along the sagittal plane (**Figure 3.6A-B**). Best frequencies increased from caudal-dorsal to rostral-ventral, along the long axis of intermediate-a. To quantify spatial progressions in BF, we used multiple linear regression to predict BFs using the rostral-caudal (RC) and dorsal-ventral (DV) coordinates. For neurons that were 0.7 mm to 1.0 mm from the midline, a linear model with RC and DV as predictor variables were significantly predictive of BFs (**Figure 3.6A, bottom; Table 3.1;** $F = 22.0$, $p < 0.001$, $R^2 = 0.47$). For neurons that were 1.0 mm to 1.3 mm from the midline, the model was likewise a significant predictor of BFs, but the percentage of variance explained was lower (**Figure 3.6B, bottom; Table 3.2;** $F = 23.3$, $p < 0.001$, $R^2 = 0.19$). For neurons that lie further than 1.3 mm from the midline, RC and DV coordinates did not predict BFs (**Figure 3.6C-D;** 1.3 mm - 1.6mm & 1.6 mm - 1.9 mm: $F = 1.8$ & 2.0, $p > 0.1$, $R^2 = 0.01$ & 0.05).

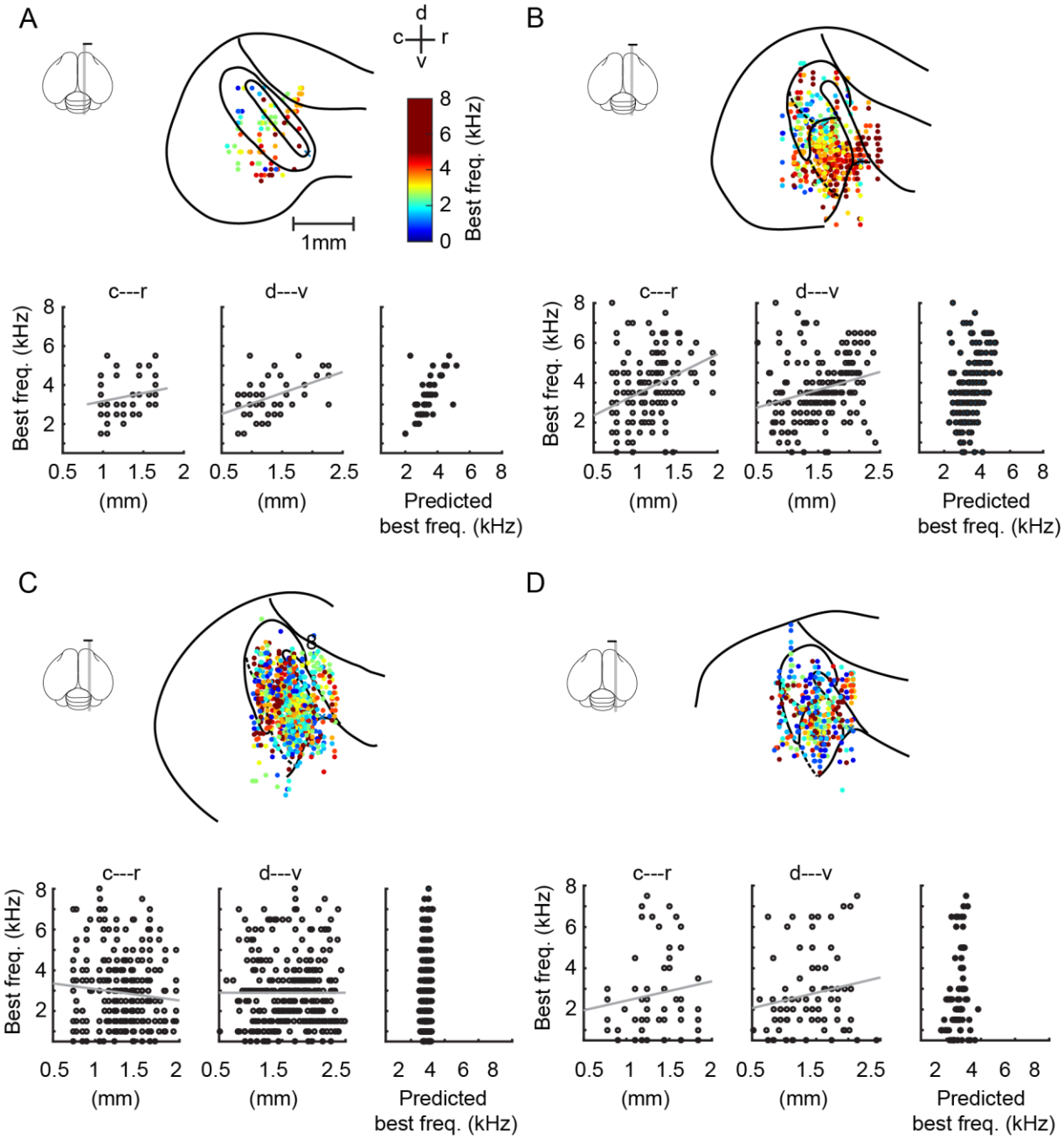


Figure 3.6. Tonotopic organization along the sagittal plane is restricted to the medial aspects of the auditory cortex.
(A, B, C, D top) show tone-responsive single units, color-coded by their best frequencies, overlaid on anatomical tracings of sagittal sections. Neurons were segmented by their distance from the midline into the following groups: *(A)* 0.7 mm to 1.0 mm, *(B)* 1.0 mm to 1.3 mm, *(C)* 1.3 mm to 1.6 mm, and *(D)* 1.6 to 1.9 mm. Shown below each anatomical plot are scatter plots showing how best frequencies vary with rostral-caudal (RC) coordinate (left) and how best frequencies vary with dorsal-ventral (DV) coordinate (middle). Gray lines show best fit least-squares regression lines. Best frequencies predicted by a multiple linear regression model with RC and DV coordinates are plotted against actual best frequencies (right).

	b	SE b	β	Sig.
Intercept	-1249.1	775.9	-0.04	
Rostral-caudal coordinate	1958.6	430.3	0.52	***
Dorsal-ventral coordinate	1551.0	246.8	0.71	***

Table 3.1. Spatial variation of best frequencies in the auditory cortex (0.7 - 1.0 mm from midline).
Table shows results of multiple linear regression using rostral-caudal and dorsal-ventral coordinates to predict best frequencies of auditory cortical neurons 0.7 to 1.0 mm from the midline. b = coefficient for the given predictor variable; SE b = standard error of coefficient; β = standardized coefficient; Sig = significance level of coefficient. *** $P < 0.001$

	b	SE b	β	Sig.
Intercept	797.9	428.8	-0.02	
Rostral-caudal coordinate	1763.2	344.7	0.34	***
Dorsal-ventral coordinate	571.7	209.6	0.19	**

Table 3.2. Spatial variation of best frequencies in the auditory cortex (1.0 - 1.3 mm from midline).
Table shows results of multiple linear regression using rostral-caudal and dorsal-ventral coordinates to predict best frequencies of auditory cortical neurons 1.0 to 1.3 mm from the midline. b = coefficient for the given predictor variable; SE b = standard error of coefficient; β = standardized coefficient; Sig = significance level of coefficient. *** $P < 0.001$, ** $P < 0.01$

3.3.6 Anatomical organization of spectral selectivity for call stimuli

To assess neuronal sensitivity to spectral degradation, we compared each neuron's responses to vocoded versions of the natural calls (e.g. **Figure 3.5D**). A spectral selectivity index (SSI) was used to compare units' responses to high-resolution vocoded calls (40 and 80 channels) and low-resolution vocoded calls (16 and 20 channels). The SSI was defined as the average firing rates evoked by high-resolution vocoded calls minus that evoked by low-resolution vocoded calls, normalized by their sum (SSI further detailed in Section 2.5.6).

To determine the spatial organization of response selectivity, we anatomically mapped SSI by plotting the locations of recorded neurons based on reconstructed recording coordinates and overlaying these plots on tracings of AC regions (**Figure 3.7B**). Maps showed that SSI was spatially organized, with neurons with positive selectivity (blue) concentrated in the region ventral and caudal to intermediate-a (**Figure 3.7B, top-right** and **bottom-left**). Neurons with positive spectral selectivity did not appear concentrated along particular iso-frequency columns along the tonotopic map; further analysis of the relationship between best frequencies and SSI is described in Section 4.3.3. The medial-lateral planes with the strongest clustering of positive spectral selectivity (1.0 mm - 1.6 mm, **Figure 3.7B,** top-right and bottom-left) traversed the planes that we identified as tonotopic along the sagittal plane (1.0 mm - 1.3 mm, **Figure 3.6B**) and the planes where we did not identify tonotopic organization along the sagittal plane (1.3 – 1.6 mm, **Figure 3.6C**).

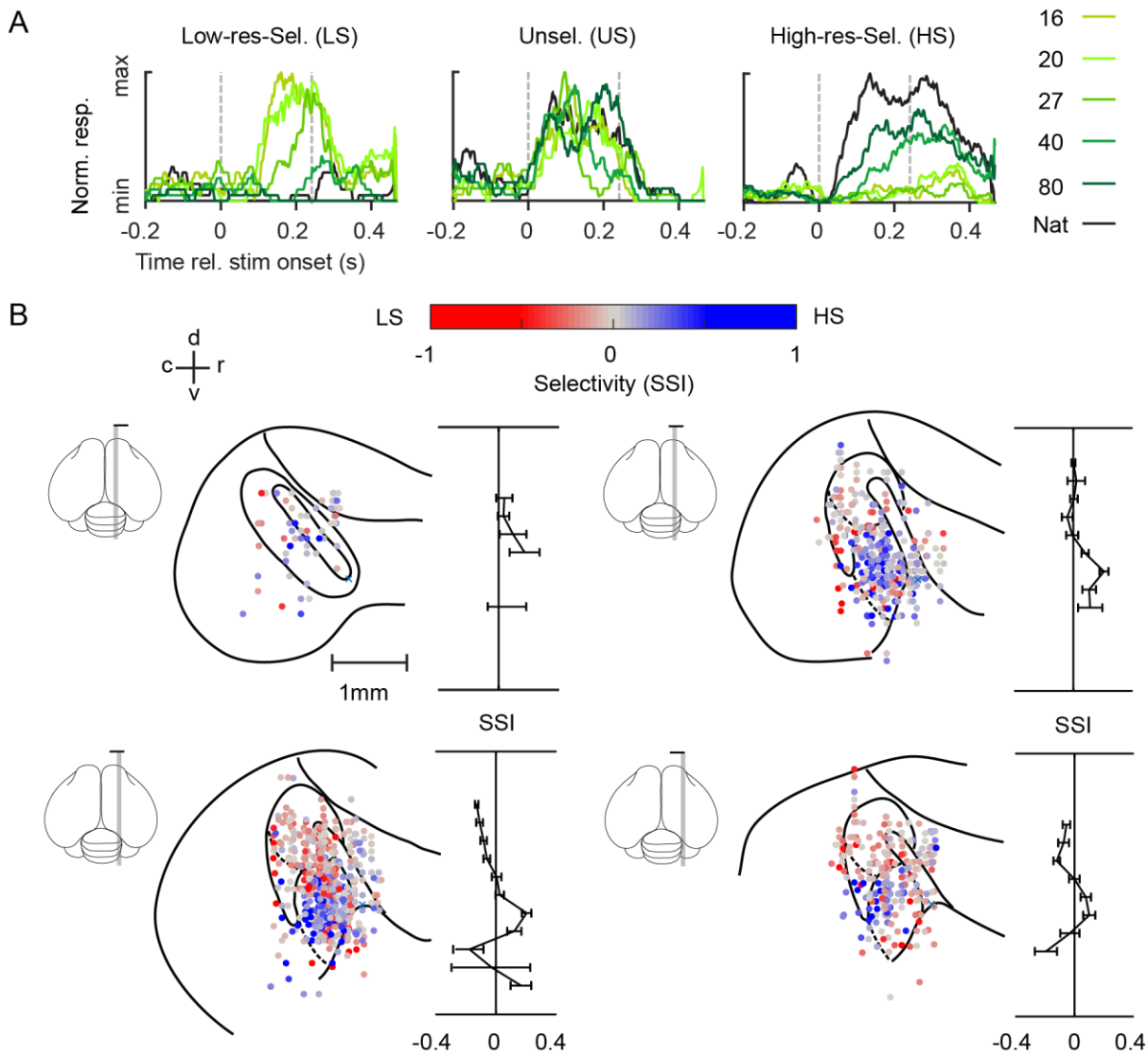


Figure 3.7 Neurons selective for high spectral resolution were localized to the deep region. *(A)* Normalized peri-stimulus time histograms (PSTHs) of three single units that were classified as low-resolution-selective (LS), unselective (US), and high-resolution-selective (HS). Responses to natural calls and their vocoded versions are overlaid and color-coded according to the rightmost legend. *(B)* Spatial organization of SSI in the AC (N = 1154 call-responsive units). Each of the four sagittal brain diagrams show single units within a 0.3mm range on the medial-lateral axis (estimated medial-lateral coordinates: 0.7 - 1.0mm, 1.0 - 1.3mm, 1.3 - 1.6mm, and 1.6 - 1.9mm), plotted according to their rostral-caudal and dorsal-ventral coordinates. Each data point represents a single unit and is color-coded according to its SSI. To the right of each brain section diagram, average SSIs (mean ± SEM) for each dorsal-ventral position bin are plotted. Note that average SSI was only computed for bins with at least 5 single units recorded.

3.3.7 Spectral selectivity differs between auditory regions, but not between putative excitatory principal cells and putative inhibitory interneurons

Given our observation that spectral selectivity is spatially organized in the auditory cortex, we compared the distribution of SSIs across anatomically defined brain regions. We also tested whether SSIs differed between the two major cell types in the auditory cortex.

As in mammals, the songbird auditory cortex is comprised of two major electrophysiological cell types that differ in action potential width and average firing rate. Neurons with narrower action potentials and higher firing rates are thought to be excitatory principal cells, and neurons with broader action potentials and lower firing rates are thought to be inhibitory interneurons (Araki et al., 2016; Harris & Mrsic-Flogel, 2013; Meliza & Margoliash, 2012). Because the correspondence between electrophysiological cell type and morphological or biochemical features has not been experimentally established in the songbird, we refer to them as putative excitatory principal cells (pPCs) and putative inhibitory interneurons (pINs).

In all four regions examined, neurons showed a bimodal distribution of action potential widths (the width at half-height of the negative peak, plus the width at half-height of the positive peak) (**Figure 3.8C,** top). Neurons were separated into two clusters based on action potential width using a Gaussian mixture model (**Figure 3.8A, C**).

We assessed whether each region's pINs and pPCs were significantly low-resolution- or high-resolution-selective by testing whether the mean SSIs were significantly different from zero (**Figure 3.8A-B**). There were no significant differences between pINs and pPCs in any region examined. In the deep region, both pINs and pPCs were selective for high-resolution calls ($p < 0.001$, one sample t-test). In the intermediate-b region, both pINs and pPCs in were selective for

low-resolution calls ($p < 0.001$, one-sample t-test). SSIs in the intermediate-a region and secondary region did not differ significantly from zero ($p > 0.5$, one-sample t-test).

Similar to previous studies, we found that neurons with greater action potential width have lower firing rates to auditory stimuli (Araki et al., 2016; Calabrese & Woolley, 2015). A significant negative correlation was found between call-evoked firing rates and action potential width in all four regions examined (**Figure 3.9C,** middle; Pearson's r between -0.57 and -0.51, all $p < 0.001$). There was no correlation between action potential width and SSI in all four regions examined (**Figure 3.9C,** bottom, Pearson's r between 0.04 to 0.16, all $p > 0.05$).

As cell types did not differ in spectral selectivities, we combined them in subsequent data analyses in this chapter. For the four regions examined, we analyzed neurons' response selectivities (proportion of stimuli that did not evoke significant responses, as in **Figure 3.5C**) for each stimulus category (**Figure 3.9**). We found that only the deep region had response selectivities that differed between stimulus categories. A greater proportion of high-resolution- than low-resolution vocoded calls evoked significant responses from deep-region neurons (**Figure 3.9C**, ANOVA, $F(6, 2870) = 9.21$, $p < 0.001$; Tukey's test, $p < 0.05$).

Taken together, these results indicate that high-spectral-resolution-selective neurons are localized to the deep region, but do not correspond to one particular auditory cortical cell type.

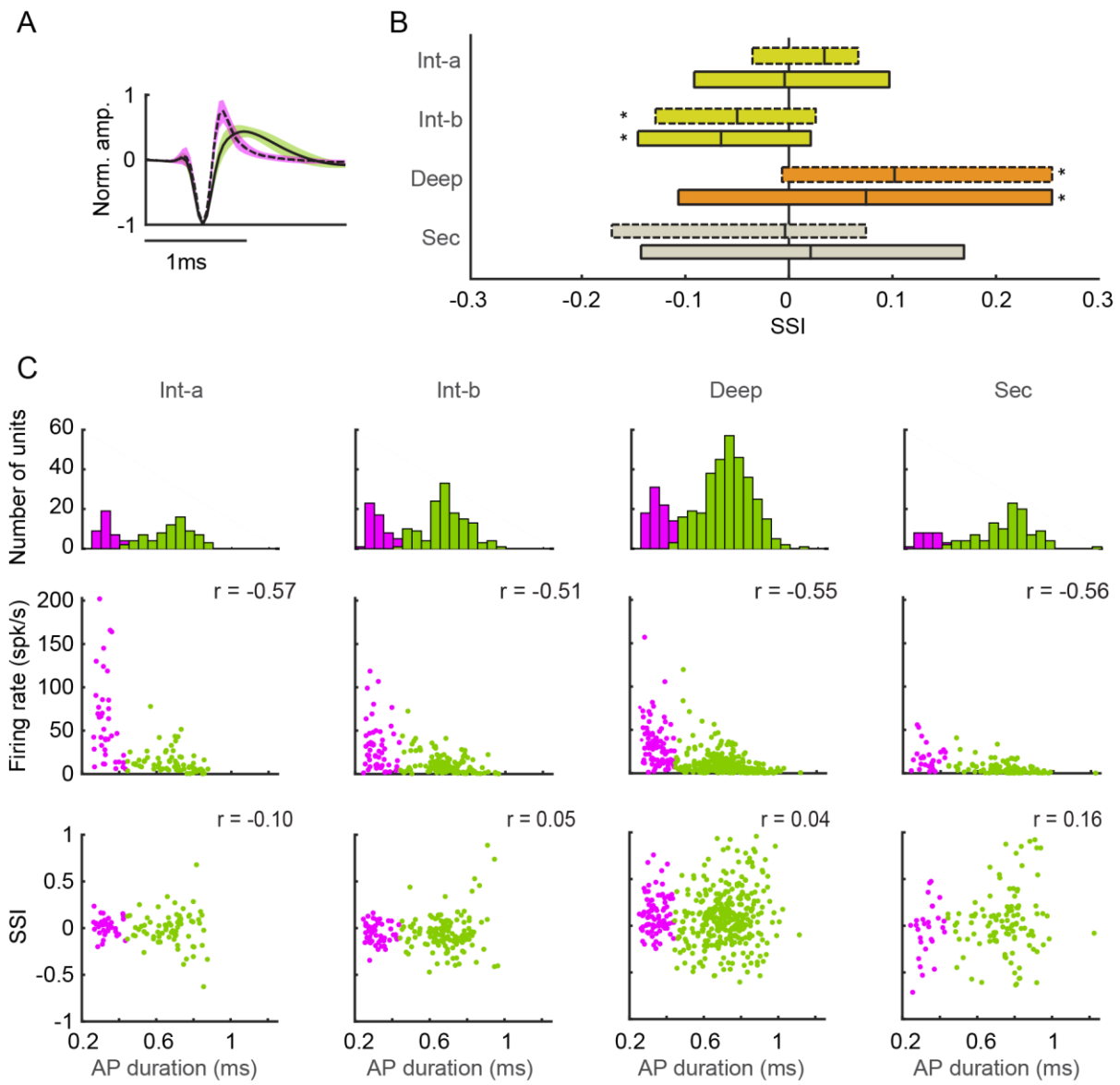


Figure 3.8 Call firing rates and spectral selectivity by cell type and auditory region. *(A)* Average action potential waveforms (mean ± SEM) of putative inhibitory interneurons (pINs, magenta) and putative excitatory principal cells (pPCs, green). Amplitudes were scaled to range between -1 and 1. *(B)* Boxplots showing the distribution of spectral selectivity indices (SSIs) for pINs and pPCs in four auditory cortex regions. Asterisks show significant differences from zero (one-sample t-test, $p < 0.001$). Deep region selectivities are significantly positive and Int-b region selectivities are significantly negative. *(C, top)* Histogram of action potential durations for each auditory region examined. *(C, middle)* Scatter plots of action potential duration against firing rate for each region examined. *(C, bottom)* Scatter plots of action potential duration against SSI for each region examined. Pearson's r correlation coefficients are shown on the top-right of each scatter plot.

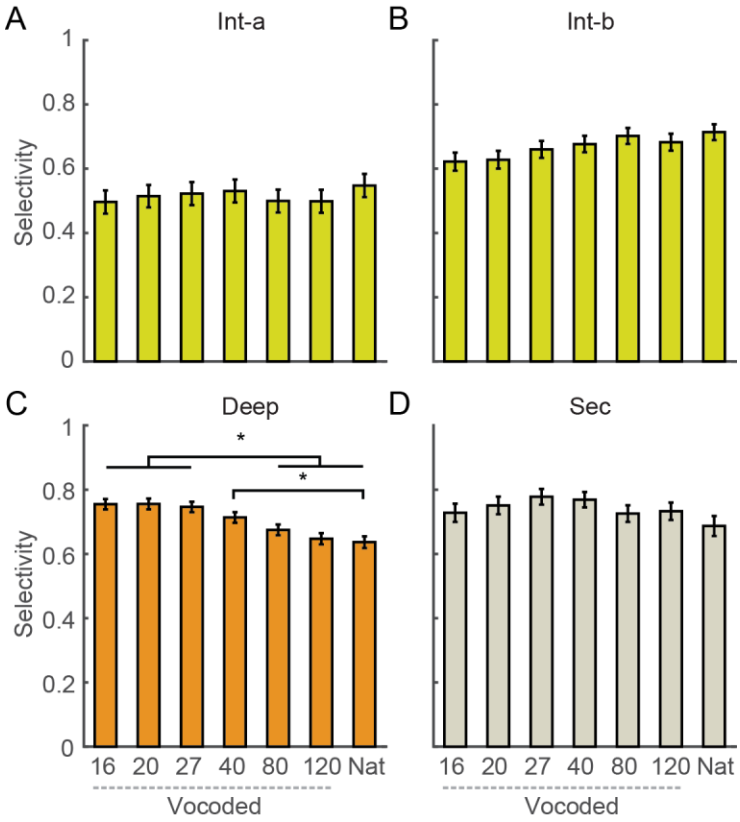


Figure 3.9 Response selectivity for call stimuli in the songbird auditory cortex. Selectivity, defined as the proportion of stimuli that did not elicit a significant response, for different stimulus categories in *(A)* intermediate-a (N = 111), *(B)* intermediate-b (N = 189), *(C)* deep (N = 411), and *(D)* secondary regions (N = 136). Bar graphs show mean ± SE. Asterisks denote significant differences revealed by Tukey tests following one-way ANOVA, $P < 0.05$.

3.3.8 Spectral selectivity is time-window dependent

To investigate deep region responses to call stimuli in more detail, we segmented neurons into three groups by their SSI and compared their spontaneous firing rates, as well as stimulus-evoked firing rates in different time windows. Neurons with SSI < -0.2, indicating 50% greater response to low-resolution vocoded calls (16 and 20 channels) were classified as low-resolution-selective (LS). Neurons with SSI > 0.2, indicating 50% greater response to high-resolution vocoded calls (40 and 80 channels), were classified as high-resolution-selective (HS). The remaining neurons with -0.2 < SSI < 0.2 were classified as unselective (US) (**Figure 3.10A**).

LS, US, and HS neurons differed significantly in spontaneous firing rates, with US neurons having higher spontaneous rates than LS and HS neurons (**Figure 3.10B,** ANOVA, $F(2, 408) = 7.34$, $p < 0.001$; Tukey's tests, $p < 0.05$). Call-evoked firing rates, taken as the highest firing rate elicited by vocoded or natural calls for each neuron, did not differ significantly between the three groups of neurons (**Figure 3.10C,** ANOVA, $F(2,408) = 2.93$, $p > 0.05$). Call-evoked response strength (call-evoked firing rate minus spontaneous rate for each neuron) did not differ between the three groups of neurons (**Figure 3.10D**, ANOVA, $F(2,408) = 1.93$, $p > 0.05$).

We investigated how neurons' responses vary with spectral resolution by plotting average response strengths of LS, NS, and HS neurons for each stimulus category (**Figure 3.10E**). We considered different time windows to calculate spike rates, and found that neurons' stimulus response curves differ by time window.

When considering a broad time window – including all spikes between stimulus onset and 20 ms after stimulus offset – we found that there was a significant main effect of neuron group on firing rates (**Figure 3.10E, left,** ANOVA, $F(2, 2856) = 27.57$, $p < 0.001$). While there was no significant main effect of stimulus type (**Figure 3.10E, left,** ANOVA, $F(6, 2856) = 0.65$, $p > 0.05$),

there was a significant interaction between selectivity group and stimulus type (**Figure 3.10E, left,** ANOVA, $F(12, 2856) = 0.65$, $p < 0.001$). This indicates that the effect of stimulus type on neural responses is dependent on the selectivity group of neurons.

During the onset period (first 50 ms after stimulus onset), US neurons had higher responses than LS and HS neurons (**Figure 3.10E, second left,** ANOVA, $F(2, 2856) = 83.08$, $p < 0.001$), but there was no significant effect of stimulus type or interaction effect between neuron group and stimulus type (both $p > 0.1$).

Response during the sustained time window (between 50ms after stimulus onset, and stimulus offset) and during the offset period (first 50 ms after stimulus offset) were similar to that of the broad time window (**Figure 3.10E, two right figures**). For both sustained and offset response strengths, there was a significant main effect of selectivity group ($p < 0.001$), no significant main effect of stimulus type ($p > 0.7$), and a significant interaction between selectivity group and stimulus type ($p < 0.01$).

These results indicate that sensitivity to spectral resolution develops after the first 50 ms of stimulus presentation. Spectral resolution sensitivity during the sustained and offset periods drives overall sensitivity.

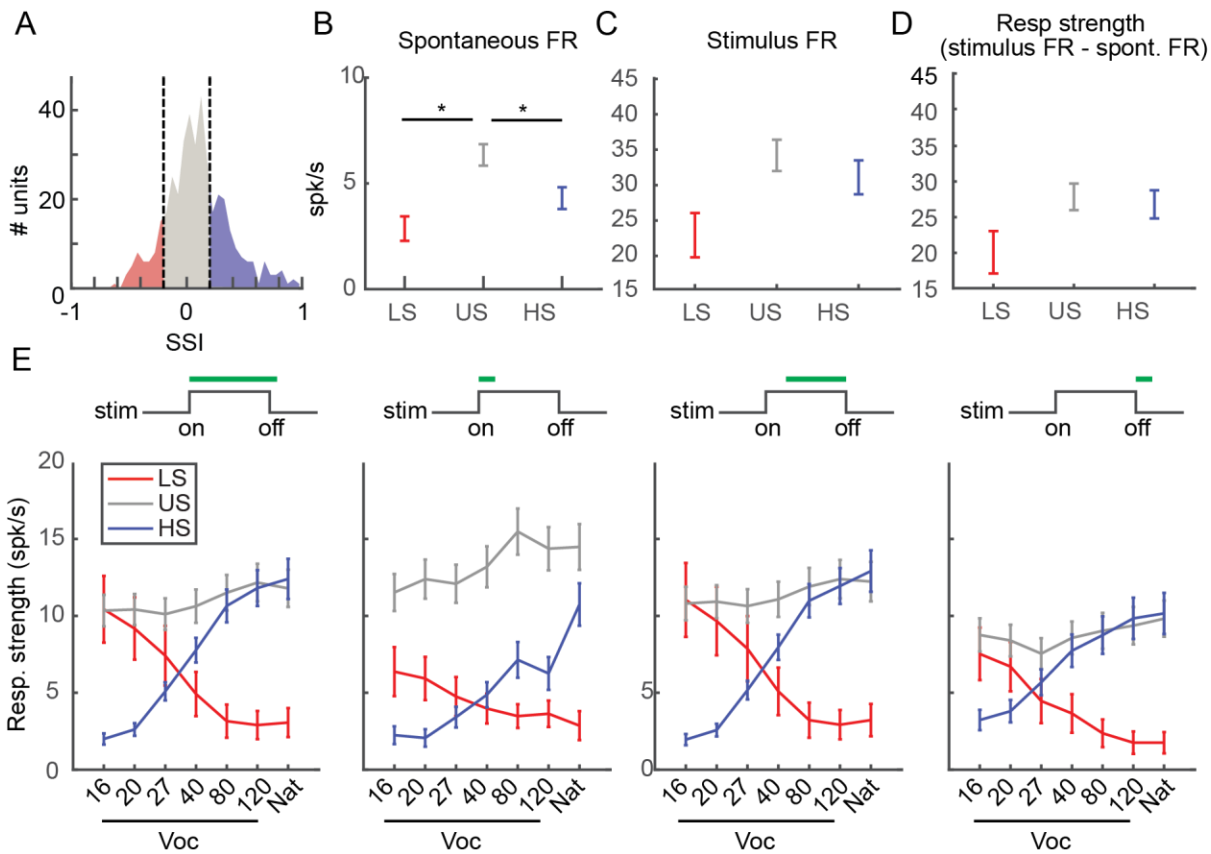


Figure 3.10 Deep region response strengths to different stimuli and at different time windows.
(A) Histogram showing distribution of spectral selectivity index (SSI) of deep region neurons. Dotted lines show SSI cutoffs for segmentation. Red, low-resolution-selective (LS); gray, unselective (US); blue, high-resolution-selective (HS). *(B, C, D)* Spontaneous firing rates *(B)*, stimulus-evoked firing rates *(C)*, and response strengths *(D)* of LS, US, and HS neurons. Asterisks indicate significant differences between groups (ANOVA and Tukey's tests, $p < 0.05$). *(E)* Response strengths of LS, US, and HS at different time windows relative to stimulus presentation. Schematics above each graph show the time window during which spike rates were quantified. Overall window (left): from stimulus onset until 20 ms after stimulus offset; onset window (second left): the 50 ms following stimulus onset; sustained window (second right): the time subsequent to the onset window until stimulus offset; offset window (right): the 50 ms after stimulus offset.

3.4 Discussion

We found that spectral resolution of communication calls is behaviorally relevant, and is represented by a distinct neuronal population within the auditory cortex. A robust neural sensitivity to spectral resolution emerges within the cortical circuit between the thalamorecipient and deep regions. Spectral selectivity of neurons in the deep output region may serve the extraction of vocalizations from the environment and the perception of social information carried in those signals.

The spectral structure of speech is important for human communication (further reviewed in Section 1.3.3). It contributes to the extraction of social information from voices, as well as the ability to understand speech in the presence of interfering signals (Gonzalez & Oliver, 2005; Popham et al., 2018). Our experiments demonstrate that for zebra finches, the spectral features of calls must be preserved to a certain degree to elicit vocal responses. To elicit responses, calls must contain distinct spectral peaks and valleys, but the spectral peaks need not be harmonically related.

Our results add to existing knowledge on acoustic features that drive social responses to vocalizations in the zebra finch. In the spectral domain, call stimuli that have F0s within the range of about 550-750 Hz (Vicario et al., 2001) evoke stronger behavioral responses. Wideband calls are preferred to narrowband calls, with at least four frequency components required to elicit typical behavioral responses (Vignal & Mathevon, 2011). Temporally, calls with longer duration (Vicario et al., 2001) and naturalistic amplitude modulation (Vignal & Mathevon, 2011) evoke stronger responses. In our study, only 40- and 80-channel vocoded calls had distinct frequency components, and only those vocoded calls evoked significant vocal responses. This selectivity is likely not a response to coarse spectral shape and amplitude envelope, as vocoded calls with differing channel

numbers have similar coarse spectral shapes and amplitude envelopes that are highly correlated to those of natural calls (see **Figure 2.1**).

We also found that birds treated inharmonic calls similarly to harmonic calls, despite their ability to detect fine frequency shifts in harmonic sounds (discrimination thresholds < 1Hz) (Lohr & Dooling, 1998) . Hence, while birds likely discriminate between inharmonic and harmonic calls, this difference may not lead to differences in vocal responses to these sounds. However, since birds can perceive the difference between inharmonic and harmonic calls, our results do not preclude the possibility that harmonicity may be important for behaviors other than vocal responses to call playbacks. For example, female preference for male song (S. C. Woolley & Doupe, 2008) may depend on the presence of harmonically related frequencies in harmonic stack syllables, and birds may preferentially elicit playbacks of harmonic vocalizations over inharmonic vocalizations. In addition, birds could use harmonicity as a grouping cue to extract individual vocalizations from a mixture of different birds' vocalizations (Schneider & Woolley, 2013). Study of the perception of inharmonic signals in a variety of behavioral contexts would aid in further understanding how birds use harmonic structure for communication.

The importance of spectral structure could differ between types of auditory tasks. A previous study showed that birds trained to recognize harmonic musical tone sequences recognized noise-vocoded versions that lack deep spectral modulations and harmonicity, as long as the coarse spectral shape was preserved (Bregman et al., 2016). In contrast, our study shows that coarse spectral shape is insufficient to elicit vocal responses. Further studies are needed to determine the significance of spectral resolution for other communication tasks such as extracting signals from sound mixtures.

In addition to tonotopic organization, which we found to be more prominent in the medial portion than in the lateral portion of auditory cortex, we also identified robust anatomical localization of spectral resolution sensitivity. Our identification of high-resolution-selective neurons in the deep region of primary auditory cortex suggests that the processing of behaviorally relevant sounds engages specialized neural pathways. Spectral resolution sensitivity in deep region neurons is not apparent in the first 50 ms after stimulus onset but emerges subsequently.

To understand the tuning properties driving spectral resolution selectivity, and to link neural responses to birds' behavioral preferences, we further characterized neural responses in the deep region using a wide array of synthetic and natural stimuli. These analyses form the basis of Chapter 4.

Chapter 4

SPECTRAL MODULATION DEPTH SENSITIVITY UNDERLYING SELECTIVITY FOR BEHAVIORALLY RELEVANT VOCALIZATIONS

4.1 ABSTRACT

Harmonicity and spectral modulation are omnipresent and often intertwined in natural vocalizations. However, they may differ in importance for communication behavior and engage fundamentally different auditory processing mechanisms. Therefore, it is important to understand which of these spectral properties underlie the neural processing and perception of vocalizations. Spectral degradation – reducing frequency information into a small number of spectral channels – affects the perception of social information from vocalizations. Our behavioral studies described in Chapter 3 showed that spectral degradation of vocalizations reduces evoked vocal responses from songbirds. Rendering vocalizations inharmonic did not impair vocal responses, indicating that birds require spectral modulations (fluctuations in power along the frequency axis), but not harmonicity (integer ratios among frequency components), for behavioral responses.

In the previous chapter, we described a distinct population of neurons in the deep region whose responses decrease with spectral degradation. Here, we further characterize the responses of these neurons to calls and a range of other stimuli to understand their response dynamics and spectral tuning properties. First, we investigated whether neurons that show spectral selectivity for calls can function as detectors of particular spectral structure in the context of song, where individual sound elements (syllables) range from noisy to spectrally modulated and harmonic. We

found that high-resolution-selective responses were robust, persisting with changes in acoustic context. By analyzing responses to broadband spectrally rippled sounds ("ripples"), we identified sensitivity to spectral modulation depth, instead of to harmonicity, as a driver for these neurons' selectivity for behaviorally relevant vocalizations. Finally, by characterizing the temporal response properties of neural responses to calls, ripples, and tones, we reveal that neurons with high spectral selectivity are characterized by distinct temporal response patterns, suggesting a difference in the inputs that they receive.

4.2 INTRODUCTION

The previous chapter (**Chapter 3**) established that birds' vocal responses to communication calls require spectral resolution to be sufficiently high to reconstruct spectral peaks and valleys. Neurons in the deep region, but not those in thalamorecipient regions, show corresponding selectivity for calls with high spectral resolution.

Here, we begin by characterizing deep region neurons' responses to song, which is a complex string of acoustic elements (syllables) that show divergent acoustics. Syllables naturally range from noisy to deeply spectrally modulated and harmonic, providing an opportunity to test whether spectral selectivity (as assessed using vocoded calls) generalizes to the processing of natural variations in spectral structure. In addition, previous studies have shown that auditory cortical neurons' responses to particular syllables can be modulated by whether they are preceded by silence or by other syllables (Schneider & Woolley, 2013), so it is possible that spectrally selective responses, which we quantified by presenting isolated sound elements, may differ when neurons are processing song. We asked whether deep region neurons' spectral selectivity would persist in the context of song, and showed that spectral selectivity is robust to acoustic context;

neurons that are selective for high-resolution calls also show elevated responses to syllables with more prominent spectral modulations and harmonic structure within songs.

Spectral selectivity as exhibited by deep region neurons requires identifying and representing specific differences in acoustic structure between high-resolution vocoded calls (40 and 80 channels) and low-resolution ones (16 and 20 channels), and between noisy syllables and harmonic stack syllables in song. Spectral degradation, achieved by reducing the number of channels in a noise vocoder, decreases both spectral modulation depth and harmonicity. Similarly, harmonic stack syllables in a zebra finch song contain deeper spectral modulations and greater harmonicity than noisy syllables. While deep spectral modulations and harmonicity typically co-occur in vocalizations, their neural processing may engage fundamentally different neural mechanisms. Our studies investigating neural responses to spectrally modulated ripples ("ripples") with varying phases and depths aimed to dissociate these two spectral features and determine how each feature contributes to spectral selectivity for behaviorally relevant communication calls.

We defined spectral modulation depth as the difference in amplitude (dB) between peaks and valleys of spectral energy. AC neurons that are sensitive to amplitude differences across frequencies have been identified in guinea pigs (Catz & Noreña, 2013) and marmosets (Barbour & Wang, 2003), and enhanced neural representation of spectral peak-to-valley differences is predicted by computational models to result from lateral inhibition (Shamma, 1985; Yost, 1986).

Harmonicity was defined as the alignment between frequency components of a stimulus with integer multiples of an F0 (harmonic template). Previous studies have shown that a sensitivity to harmonic structure can result from non-linear facilitation of neural responses to combinations of harmonically related tones (Feng & Wang, 2017). The detection of harmonically related frequency components is a proposed mechanism in spectral models of pitch extraction (Duifhuis

et al., 1982; Scheffers, 1983). Through analyzing deep region population responses to ripples and relating ripple responses to calls spectral selectivity, we established that selectivity for behaviorally relevant calls is driven by sensitivity to modulation depth, and not to harmonicity.

In addition to identifying spectral features driving selectivity for communication calls, we also characterized how neural responses developed over time. We measured first-spike latencies to tones, ripples, and calls, and the temporal dynamics of neural responses to call stimuli. Latencies of response can be an indicator of the neurons' hierarchical position along the processing pathway, and may encode stimulus features (Kikuchi et al., 2014; Mormann et al., 2008; Phillips, 2000). Neural responses at different time windows have been shown to differ in stimulus selectivity, with onset firing playing different roles than sustained firing in stimulus representation (X. Wang et al., 2005).

4.3 RESULTS

4.3.1 Song representation transforms along the auditory cortical pathway

In order to investigate how neurons in the auditory cortex encode complex vocalizations in which acoustic features naturally vary over time, we recorded auditory cortical neurons' responses to song. For a subset of neurons that we collected call response data from, we also presented five zebra finch songs as stimuli (e. g. **Figure 4.1**).

We examined the pPSTHs of intermediate-a, intermediate-b, deep, and secondary regions evoked by song stimuli. Call-responsive neurons in all regions examined showed population activity that fluctuated during song presentation (**Figure 4.2**). Average firing rates to song differed significantly between regions and between putative principal cells (pPCs) and putative inhibitory interneurons (pINs) (**Figure 4.3A;** ANOVA, $F_{region}(3, 395) = 43.2$, $p < 0.001$; $F_{cell\ type}(1, 395) = 119.0$, $p < 0.001$). Specifically, intermediate-a pPCs had stronger responses to song than pPCs in

all other examined regions, and intermediate-a pINs had stronger responses to song than pINs in all other examined regions (Tukey's tests, all $p < 0.001$). pINs showed greater firing rates to song than pPCs in all brain regions (Tukey's tests, all $p < 0.001$).

Intermediate-a population responses showed sharp increases and decreases corresponding to syllable onsets and offsets. Intermediate-b, deep, and secondary region population responses were less temporally precise; responses to the edges of syllables were relatively "softened" in these regions compared to those in intermediate-a (e.g. **Figure 4.2**). We quantified each region's pIN and pPC population activity rise times (how fast the pPSTH peaks after stimulus onset) and fall times (how fast the pPSTH reaches a trough after stimulus offset) for each syllable onset and offset across the five songs. Rise times differed between brain regions (**Figure 4.3B;** ANOVA, $F(3, 516) = 14.4$, $p < 0.001$). The intermediate-a pPC population had shorter rise times than those in all other regions (Tukey's tests, $p < 0.01$), and the intermediate-a pIN population had shorter rise times than those in the deep and secondary regions (Tukey's tests, $p < 0.05$). There were no significant differences in rise times between pPCs and pINs within any region. Fall times differed between brain regions and cell types (**Figure 4.3C;** ANOVA, $F_{region}(3, 465) = 9.95$, $p < 0.001$; $F_{cell\ type}(1, 465) = 10.2$, $p < 0.01$). In intermediate-a and -b, fall times of the pIN population were significantly shorter than those of the pPC population (Tukey's tests, $p < 0.05$). Further, intermediate-a pIN population fall times were significantly shorter than those in the deep and secondary regions (Tukey's tests, $p < 0.01$). Taken together, these results indicate that population responses are more strongly time-locked to syllable onsets in intermediate-a than all other regions. Population responses to syllable offsets were cell-type specific; intermediate-a and -b pIN population responses decayed faster with stimulus offsets compared to pPC responses within regions, and to pIN responses in the secondary region.

Our findings on population responses in different auditory regions are largely consistent with previous findings. A previous study found that individual intermediate-a neurons show synchronous responses to song, leading to population responses that were deeply modulated in time (Lim et al., 2016). Individual neurons in higher regions, compared to those in intermediate-a, were shown to have more diverse responses to song, such that the population-averaged responses were relatively weakly modulated over time (Lim et al., 2016). Our data show that intermediate-a population responses were stronger and more rapidly modulated than that of other regions by syllable onsets and offsets, consistent with these previous results. Secondary region neurons have been shown to represent song with a sparse and distributed code, such that each neuron respond with spikes during narrow segments of song, collectively covering the entire song (Schneider & Woolley, 2013). Low average responses to song in the secondary region observed from our data could result from a low fraction of the population being active at any particular moment during song presentation. Population coding, as assessed by the spike timing correlation between pairs of neurons, was found to differ between brain regions and cell types (Calabrese & Woolley, 2015). Stimulus-evoked and spontaneous correlations, thought to result from shared input and direct synaptic connections respectively, are greater in pINs compared to pPCs, and greater in the deep region compared to the intermediate regions (note that intermediate-a and intermediate-b units were analyzed in combination in this previous study). Our findings show that population response dynamics to song, as assessed by the rise and decay times of neural responses after syllable onsets and offsets, also differ between cell types and between deep and intermediate-a regions.

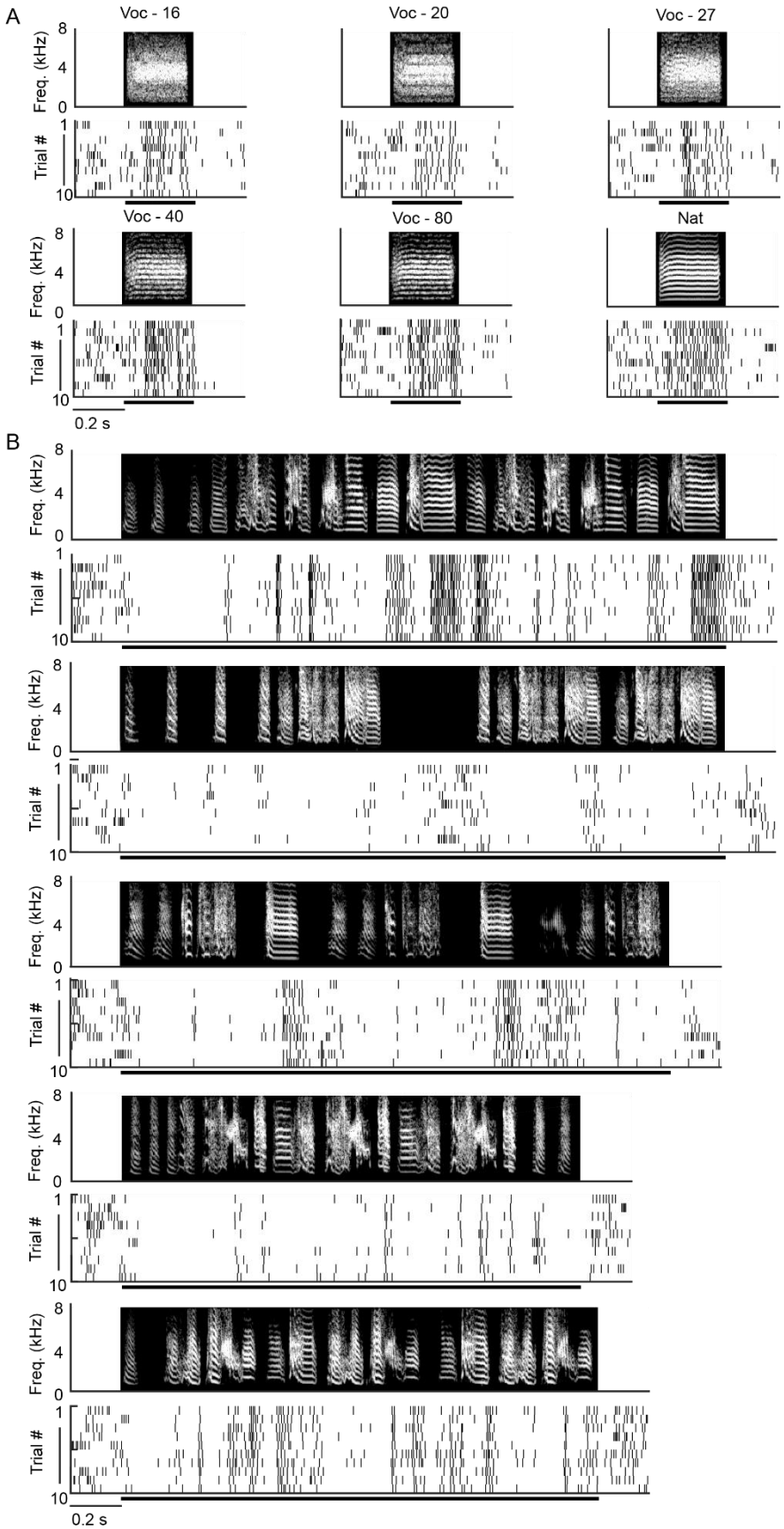


Figure 4.1 Single neuron responses to call and song stimuli. Shown example is a HS neuron from the deep region. *(A)* Raster plots showing spikes fired before, during and after the presentation of vocoded (16 to 80 channels) and natural calls. Responses are shown over 10 repetitions of each stimulus. Black bars below raster plots indicate duration of stimulus. *(B)* Raster plots showing spikes fired before, during and after the presentation of five songs. Responses are shown over 10 repetitions of each song. Black bars below raster plots indicate duration of stimulus.

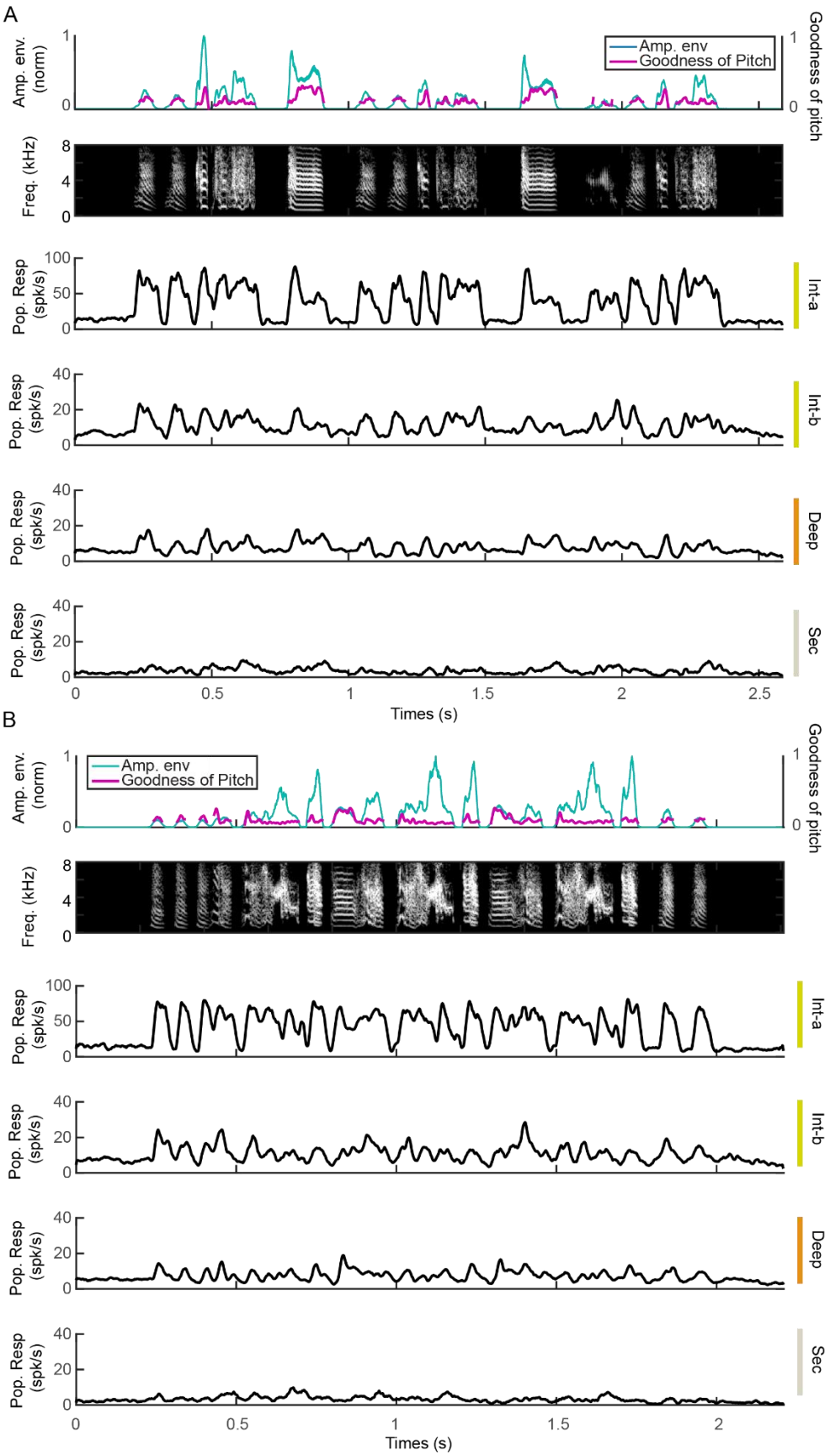

A
Amp. env (norm)
Goodness of pitch
Amp. env
Goodness of Pitch
Freq. (kHz)
Pop. Resp (spk/s)
Int-a
Int-b
Deep
Sec
Times (s)
B
Amp. env (norm)
Goodness of pitch
Amp. env
Goodness of Pitch
Freq. (kHz)
Pop. Resp (spk/s)
Int-a
Int-b
Deep
Sec
Times (s)

Figure 4.2 Population responses to song in auditory cortical regions.
(A, B) Population PSTHs of intermediate-a (N = 38), intermediate-b (N = 95), deep (N = 210), and secondary (N = 57) regions in response to two out of five presented songs. Spectrograms of song stimuli are shown directly above intermediate-a pPSTHs. Normalized amplitude envelope and feature vector for goodness of pitch are shown above spectrograms. Note that the Y-axis range for intermediate-a population responses differs from the ranges for all other regions.

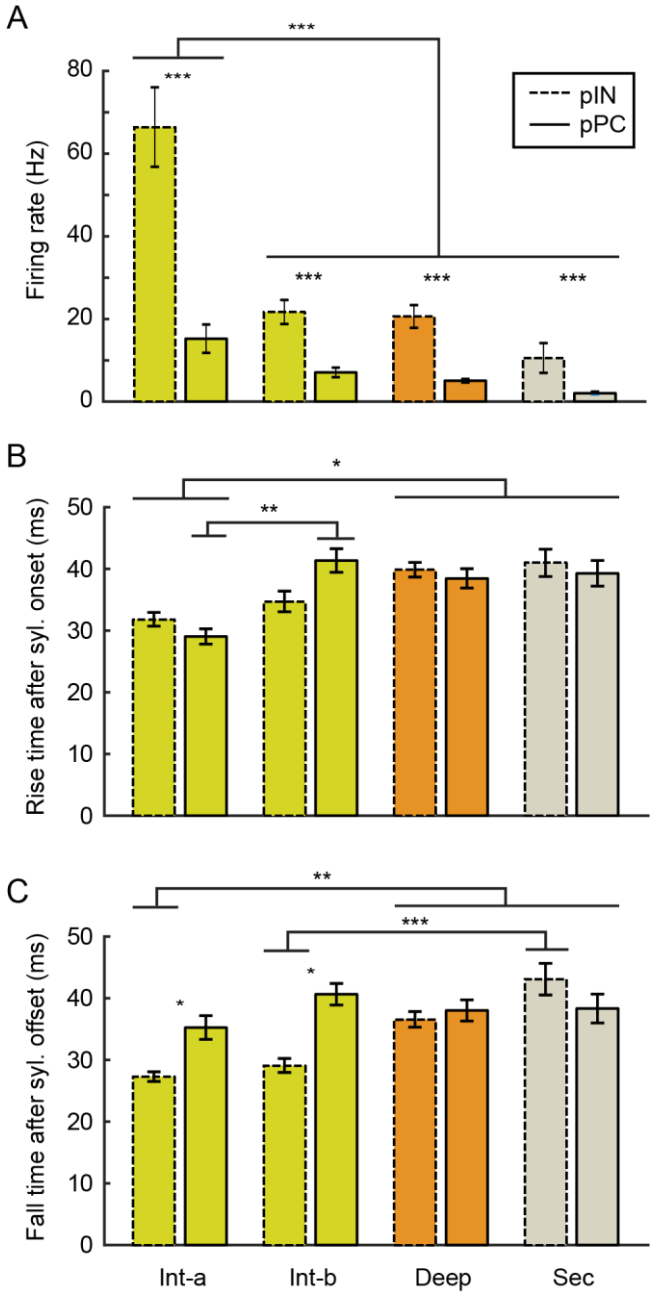


Figure 4.3 Song response magnitude and dynamics across brain regions and cell types.
(A) Bar plots (mean ± SE) showing average song-evoked firing rates of pINs (dashed outlines) and pPCs (solid outlines) in different auditory cortical regions. (N [pIN, pPC] = [21, 17], [31, 64], [44, 166], [12, 45] for intermediate-a, intermediate-b, deep, and secondary regions) *(B)* Bar plots (mean ± SEM) showing average time to peak population response after syllable onsets (N = 67 syllables). *(C)* Bar plots (mean ± SEM) showing average time to trough in population response after syllable offsets (N = 67 syllables). For all panels, asterisks directly above pairs of bars indicate significant differences between pINs and pPCs in the same brain region (Tukey's tests). Bar brackets extending over multiple bars indicate significant differences in pairwise comparisons between the same cell type in different brain regions (Tukey's tests). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Note that when multiple significant comparisons are shown with the same bar brackets, the lowest significance level is displayed.

4.3.2 Spectral selectivity to calls persists in the context of song

After characterizing the transformation of song responses across auditory regions, we examined population responses in the deep region and how they differed between neuron groups with differing call spectral selectivity. pPSTHs for LS (N = 27), US (N = 122), and HS neurons (N = 122) were constructed for each song and two examples are shown in **Figure 4.4.**

LS, US, and HS neurons' response patterns differed throughout the presentation of song stimuli. US neurons responded significantly above spontaneous firing ($p < 0.05$) during larger proportions of total song duration than did HS and LS neurons (**Figure 4.4C;** ANOVA, $F(2, 207) = 7.55$, $p < 0.001$, Tukey's tests, $p < 0.01$ for US vs. LS and US vs. HS). Observations from the group pPSTHs suggested that at times when response diverged among the three groups, US responses were higher than LS and HS responses. At other times, US responses largely overlapped with HS or LS responses.

In order to determine whether response divergences and convergences were driven by specific acoustic features in song elements, we segmented the non-silent portions of song into parts where US responses significantly exceeded HS responses (US > HS; $p < 0.05$ in two-sample, one-sided t -test) and parts where US responses were not significantly different from HS responses (US = HS; $p > 0.05$ in two-sample, one-sided t-test), and compared the acoustic features of song elements during these response portions (**Figure 4.5**). For this analysis, pPSTHs were shifted in time according to the time lag of maximum correlation between pPSTHs and the amplitude envelope of each song, in order to account for the latency of neural responses. We did not further analyze LS responses in this manner as the sample size was small (N = 27) and because LS neurons constitute the smallest portion of deep region neurons.

The song segment analysis revealed that response divergence and convergence between HS and US neurons are correlated with variations of multiple acoustic features in song. The song segments where US > HS significantly differed from the song segments where US = HS in frequency modulation, goodness of pitch, and entropy (**Figure 4.5D;** two sample t-tests, all $p < 0.001$). Specifically, US responses and HS responses were comparable during segments of song with lower frequency modulation, higher goodness of pitch, and more negative Wiener entropy. These are features characterizing the "harmonic stack" syllables in song. US responses were significantly higher than HS responses during portions of song with greater frequency modulation, lower goodness of pitch, and less negative entropy. Mean frequency did not differ between periods of divergence and convergence between HS and US neurons (**Figure 4.5D;** two-sample t-test, $p > 0.05$).

These results show that HS neurons maintain similar spectral selectivity when processing songs as that seen for the processing of vocoded calls. HS neurons preferentially respond to song elements characterized by low frequency variations over time, strongly harmonic frequency structure, and deeply modulated spectra, showing responses that are comparable in magnitude to those of US neurons. During the presentation of syllables that do not match this acoustic profile, HS responses were suppressed relative to US responses.

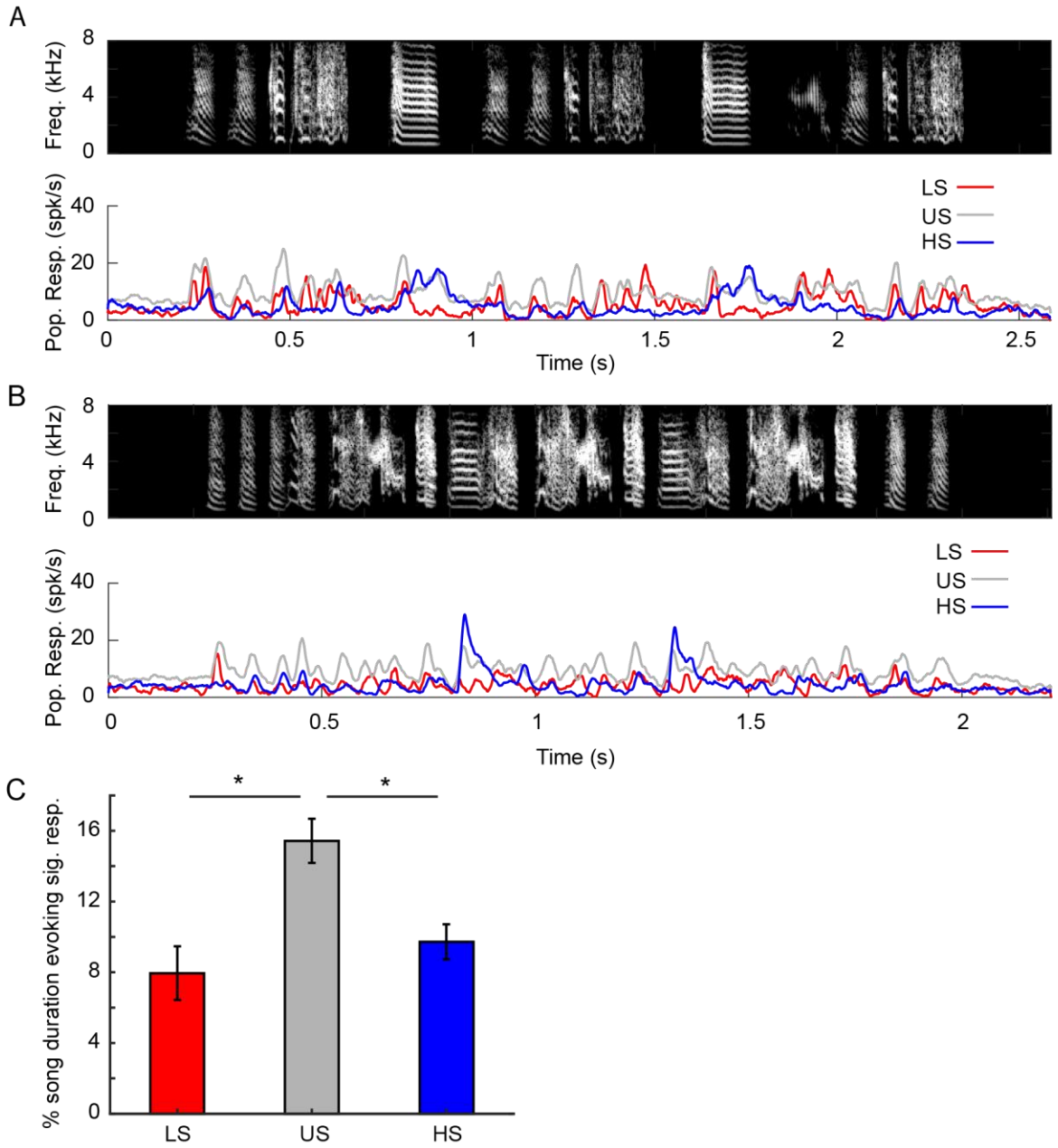


Figure 4.4 Deep region population responses to songs by call spectral selectivity group. *(A, B)* Population PSTHs of LS (N = 27), US (N = 122), and HS neurons (N = 61) to two out of five songs presented. Spectrograms are shown above pPSTHs. Shown songs correspond with those in **Figure. 4.2**. *(C)* Bar graphs (mean ± SEM) showing the average percentage of the total duration across 5 song stimuli where neurons showed responses significantly above spontaneous rates at $p < 0.05$. Asterisks indicate significant differences between LS, US, and HS neurons (Tukey's tests, $p < 0.01$).

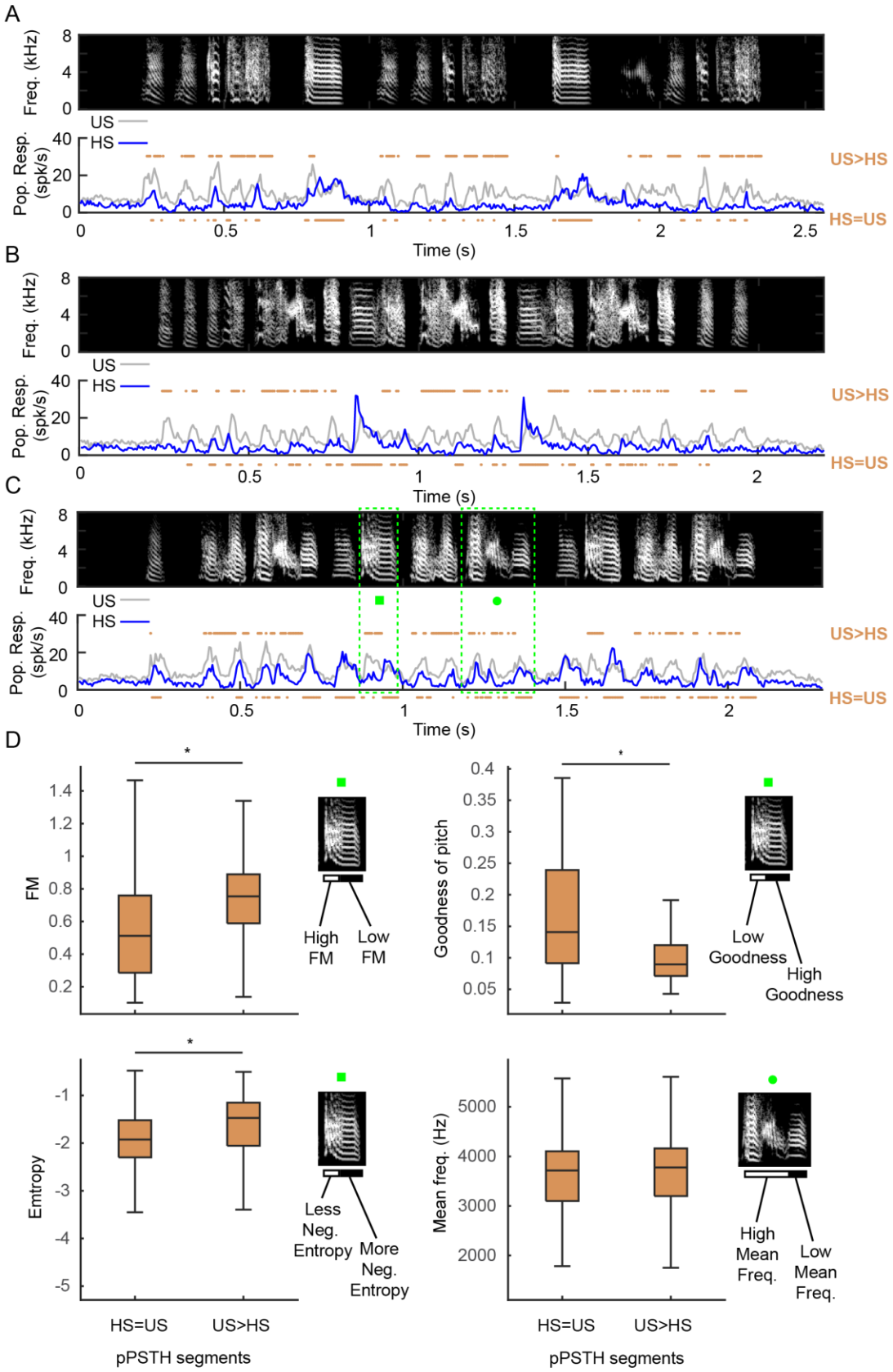

A
Freq. (kHz)
8
4
0
US
HS
Pop. Resp. (spk/s)
40
20
0
US>HS
HS=US
0
0.5
1
1.5
2
2.5
Time (s)
B
Freq. (kHz)
8
4
0
US
HS
Pop. Resp. (spk/s)
40
20
0
US>HS
HS=US
0
0.5
1
1.5
2
Time (s)
C
Freq. (kHz)
8
4
0
US
HS
Pop. Resp. (spk/s)
40
20
0
US>HS
HS=US
0
0.5
1
1.5
2
Time (s)
D
FM
1.4
1.2
1
0.8
0.6
0.4
0.2
*
High FM
Low FM
Goodness of pitch
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
*
Low Goodness
High Goodness
Emtropy
-1
-2
-3
-4
-5
*
Less Neg. Entropy
More Neg. Entropy
Mean freq. (Hz)
5000
4000
3000
2000
High Mean Freq.
Low Mean Freq.
HS=US
US>HS
pPSTH segments
HS=US
US>HS
pPSTH segments

Figure 4.5 Deep region population responses diverge and converge at acoustically distinct song segments.
(A, B, C) pPSTHs of HS and US neurons to three out of five song stimuli, with song spectrograms shown directly above pPSTHs. pPSTHs were shifted in time according to the time lag of maximum correlation between pPSTHs and the amplitude envelope of each song. Tan lines above pPSTHs indicate non-silent segments of song (5 ms bins) where US responses were significantly higher than HS responses ($p < 0.05$, two-sample, one-sided t-tests). Tan lines below pPSTHs denote non-silent segments (5 ms bins) of song where US responses were not significantly different than HS responses. Green dashed outlines and symbols in *(C)* denote syllables that are used to illustrate acoustic features in *(D). (D)* Box plots showing differences in frequency modulation (FM), goodness of pitch, Wiener entropy, and mean frequency between song segments (across all five song stimuli) where HS and US responses did not differ (HS = US), and song segments where US responses were significantly higher than HS responses (US > HS). Example song syllables illustrating variations in each acoustic feature are shown to the right of each box plot. Asterisks indicate significant differences in acoustic features between HS = US and HS > US pPSTH segments (t-tests). * $p < 0.001$; FM = frequency modulation.

4.3.3 Selectivity for call spectral structure is explained by sensitivity to spectral modulation depth

HS neurons, which responded more to high-resolution calls than low-resolution calls, also showed a sensitivity to Wiener entropy and goodness of pitch in song. Wiener entropy, which measures how uniform the sound spectrum is, becomes more negative when syllables vary from noise-like to harmonic, because harmonic stacks contain deeper spectral modulations. Goodness of pitch, which measures the strength of harmonic structure, also varies between noise-like and harmonic syllables, taking higher values for syllables with stronger harmonic structure. Hence, entropy (related to spectral modulation depth) and goodness of pitch (related to harmonicity) are naturally correlated in song, even though they measure two different aspects of spectral structure.

We analyzed responses to spectrally rippled sounds (Shamma et al., 1995), where spectral modulation depth and harmonicity were varied separately. Spectral ripples (auditory equivalent of visual gratings) are composed of sinusoidal modulations along the frequency axis. In our ripple stimuli, we varied spectral modulation depth and phase parametrically, and included spectral modulation densities of 1.2, 1.6, and 2.0 cyc/kHz (see Section 2.3.2). As shown in **Figure 4.6A,** we defined the phase at which frequency peaks aligned with integer multiples of an F0 as zero. Phase-shifted ripples were defined by the amount of shift (in proportion of a cycle) relative to the phase-aligned ripple.

Harmonicity is highest for the phase-aligned ripple (phase = zero) with maximum modulation depth, and decreases as phase deviates from zero and as spectral modulation becomes shallower (**Figure 4.6B, left**). Modulation depth specifies the peak-to-valley distance in the acoustic spectrum, and remains constant with variations in phase (**Figure 4.6B, middle**). Methods

for quantifying harmonicity by assessing stimulus alignment with a harmonic template are detailed in Section 2.5.9**.**

Deep region population responses to ripples. To obtain an overview of deep region responses to ripples, we constructed depth-phase matrices of population firing rates. Population firing rates were higher in response to ripples with larger modulation depth, irrespective of phase (**Figure 4.6B, right**)**.** To determine the respective contributions of spectral modulation depth and harmonicity, we tested whether adding each one as a predictor variable would significantly improve the ability of a multiple linear regression model to predict population firing rates (See Section 2.5.9). Harmonicity was not a significant factor in predicting population responses (Partial F-test, $F(1, 116) = 0.19$, $P = 0.66$; **Table 4.1** and **Figure 4.6C, left**); modulation depth was a significant factor in predicting population responses (Partial F-test, $F(1,116) = 136.62$, $P < 0.001$; **Table 4.2** and **Figure 4.6C, right**). By comparing the additional predictive value conferred by the inclusion of each factor, we concluded that population responses were primarily sensitive to modulation depth and insensitive to harmonicity.

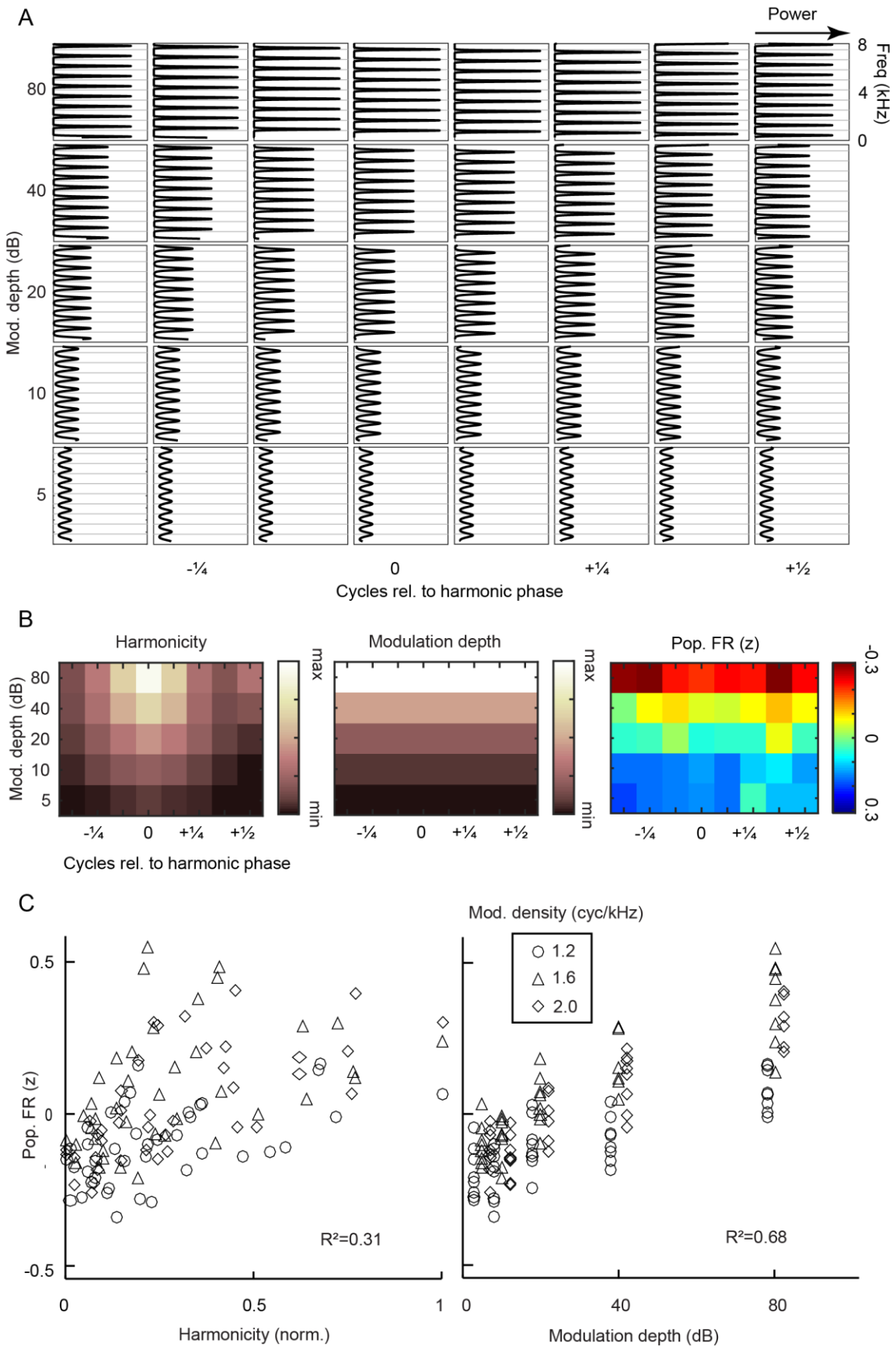

A
Power
Freq (kHz)
8
4
0
Mod. depth (dB)
80
40
20
10
5
-¼
0
+¼
+½
Cycles rel. to harmonic phase
B
Harmonicity
Modulation depth
Pop. FR (z)
max
min
-0.3
0
0.3
Cycles rel. to harmonic phase
C
Mod. density (cyc/kHz)
1.2
1.6
2.0
Pop. FR (z)
0.5
0
-0.5
R²=0.31
R²=0.68
0
0.5
1
40
80
Harmonicity (norm.)
Modulation depth (dB)

Figure 4.6 Deep region population responses to ripples are predicted by spectral modulation depth and not by harmonicity.
(A) Spectral profiles of ripple stimuli with varying phases and modulation depth. Shown spectral profiles are taken from ripple stimuli with spectral modulation density of 1.2 cyc/kHz. Gray lines indicate the expected locations of frequency components if they were integer multiples of an F0 that matches the spacing between components (F0 = 1 kHz / 1.2 = 833 Hz). *(B)* Heat maps showing how spectral modulation depth (left) and harmonicity (middle) of ripple stimuli vary with the depth and phase parameters. *(B; right)* Heat map of population firing rates in response to ripples with varying phase and depth. Each pixel represents the mean z-scored firing rate across all deep region units (N = 315 ripple-responsive units). *(C)*. Scatter plots showing how population firing rates scale with harmonicity (left) and spectral modulation depth (right). R^2 values of a linear model incorporating modulation density plus either harmonicity (left) or modulation depth (right) are shown.

	b	SE b	β	Sig.
Step 1				
Constant	-0.42	0.05		
Modulation density (cyc/kHz)	0.16	0.03	0.28	***
Modulation depth (dB)	0.0054	0.00036	0.78	***
Step 2				
Constant	-0.42	0.05		
Modulation density (cyc/kHz)	0.16	0.03	0.28	***
Modulation depth (dB)	0.0055	0.00047	0.8	***
Harmonicity	**-0.024**	**0.055**	**-0.03**	

Table 4.1 The effect of harmonicity on model prediction of deep region population responses to ripples.
Table shows multiple linear regression results for reduced model (Step 1, omitting harmonicity) and full model (Step 2, adding harmonicity). Adding harmonicity did not significantly increase in the R^2 (0.68 to 0.69; $\Delta R^2 = 0.01$; $P = 0.66$ for difference between reduced and full models). b = coefficient for the given predictor variable; SE b = standard error of coefficient; β = standardized coefficient; Sig = significance level of coefficient. *** $P < 0.001$

	b	SE b	β	Sig
Step 1				
Constant	-0.33	0.073		
Modulation density (cyc/kHz)	0.14	0.044	0.24	**
Harmonic template match	0.39	0.061	0.49	***
Step 2				
Constant	-0.42	0.05		
Modulation density (cyc/kHz)	0.16	0.03	0.28	***
Harmonic template match	-0.024	0.055	-0.03	
Modulation depth (dB)	**0.0055**	**0.00047**	**0.8**	*****

Table 4.2 The effect of modulation depth on model prediction of deep region population responses to ripples.
Table shows multiple linear regression results for reduced model (Step 1, omitting modulation depth) and full model (Step 2, adding modulation depth). Adding modulation depth significantly increased the R^2 (0.31 to 0.69; $\Delta R^2 = 0.38$, $P < 0.001$ for difference between reduced and full models). Value labels are as in Table 1. ** $P < 0.01$; *** $P < 0.001$

Relating single neurons' ripple responses to spectral selectivity for calls. To relate spectral tuning parameters to selectivity for behaviorally relevant vocalizations, we compared the ripple response properties of LS, US, and HS neurons. Example single neuron ripple depth-phase response matrices and raster plots showing responses to vocoded and natural calls are shown in **Figure 4.7.**

To determine whether deep region neurons preferred sounds with high harmonicity, we looked at their distribution of preferred phases. The best phase distributions of LS, US, and HS neurons did not differ significantly from an even distribution (**Fig 4.8A-C, left;** Pearson's chi-squared test, $p > 0.05$ for all neuron groups and all modulation densities); we did not find that neurons preferred ripples with harmonically aligned phase. The phase-response curves of LS, US, and HS neurons likewise do not indicate that neurons show maximal responses to the harmonically aligned phase (**Fig 4.8A-C, right).**

We then analyzed tuning to modulation density and depth in LS, US, and HS neurons. **Fig 4.9A** shows the depth-phase matrices for each group for the three modulation densities examined. Each neuron's matrix was centered such that the phase evoking the highest firing rate (best phase) was treated as zero. We also examined ripple response curves of LS, US, and HS neurons (**Figure 4.9B**), where firing rates to ripples with a given modulation density were normalized to range from 0 to 1 within each neuron. For LS neurons, phase modulation of responses was most apparent at 1.2 cyc/kHz. Responses generally decreased with increasing modulation depth (**Figure 4.9B, left**). For US neurons, phase and depth modulation of responses were relatively weak (**Figure 4.9B, middle**). For HS neurons, responses were higher to deeper modulations and phase modulation of responses was apparent at 1.6 cyc/kHz and 2.0 cyc/kHz (**Figure 4.9B, right**).

LS, US, and HS units showed distinct tuning profiles to modulation density. LS neurons largely preferred ripples with the lowest modulation density of 1.2 cyc/kHz (69.7%; **Fig 4.10A,** left). US neurons showed a relatively even distribution of preferred modulation density (**Figure 4.10A**, middle). HS neurons predominantly preferred ripples with higher modulation densities of 1.6 or 2.0 cyc/kHz (58.8% and 31.3% respectively; **Figure 4.10A,** right).

We examined modulation depth sensitivity in isolation by plotting depth response curves (firing rate by depth at a neurons' best phase for a given modulation density, normalized to range from 0 to 1). LS neurons decreased their responses with increasing modulation depth from 5 dB to 20 dB, and responses remained relatively constant from 20 dB to 80 dB (**Figure 4.10B, left**). US neurons increased their responses monotonically with modulation depth (**Figure 4.10B, middle**), though the increase was more moderate than for HS neurons (**Figure 4.10B, right**). Sensitivity for modulation depth was a strong predictor of SSI. Modulation depth dependency (Spearman's ρ) was used to quantify the monotonicity of the association between firing rate and modulation depth at a neuron's preferred phase for a given modulation density. Positive values indicate that firing rates increase with depth, while negative values indicate that firing rates decrease with increases in depth. HS neurons had significantly more strongly positive ρ than LS and US neurons for all spectral modulation densities examined (**Figure 4.10C**; ANOVA, $F_{1.2\ cyc/kHz}(2, 278) = 22.0$, $P < 0.001$; $F_{1.6\ cyc/kHz}(2, 278) = 50.1$, $P < 0.001$; $F_{2.0\ cyc/kHz}(2, 278) = 44.2$, $P < 0.001$; Tukey's tests, $P < 0.001$ for all pairwise comparisons and all modulation densities).

Because previous studies have shown a relationship between broader frequency tuning and lower preferred modulation density (Shamma et al., 1995), we tested whether SSI was related to frequency tuning bandwidth. We found a weak but significant negative relationship between the two when bandwidth was measured at 60 dB SPL and 70 dB SPL (**Figure 4.11A-B**, Pearson's r =

0.14 & 0.15, both $p < 0.01$), but not when bandwidth was measured at lower sound levels (**Figure 4.11C-E, -** 0.06 < Pearson's $r < 0.03$, all $p > 0.05$). SSI was unrelated to best frequencies of neurons (**Figure 4.11F**, Pearson's $r = 0.01$, $p > 0.05$).

Modulation depth dependency was significantly correlated with call SSI at all modulation densities examined (**Figure 4.12;** Pearson correlations, $r_{1.2\ cyc/kHz} = 0.35$, $p < 0.001$; $r_{1.6\ cyc/kHz} = 0.56$, $p < 0.001$; $r_{2.0\ cyc/kHz} = 0.50$, $p < 0.001$). For all three modulation densities, a majority of units (68.3%, 71.2%, and 74.0%) had modulation depth dependencies that matched the sign of selectivity for vocoded calls (i.e. positive ρ and positive SSI, or negative ρ and negative SSI).

Based on the strong preference for high modulation depth characterizing HS neurons, and the close relationship between modulation depth dependency and SSI, our results suggest that sensitivity to modulation depth underlies selectivity for the spectral structure of communication calls.

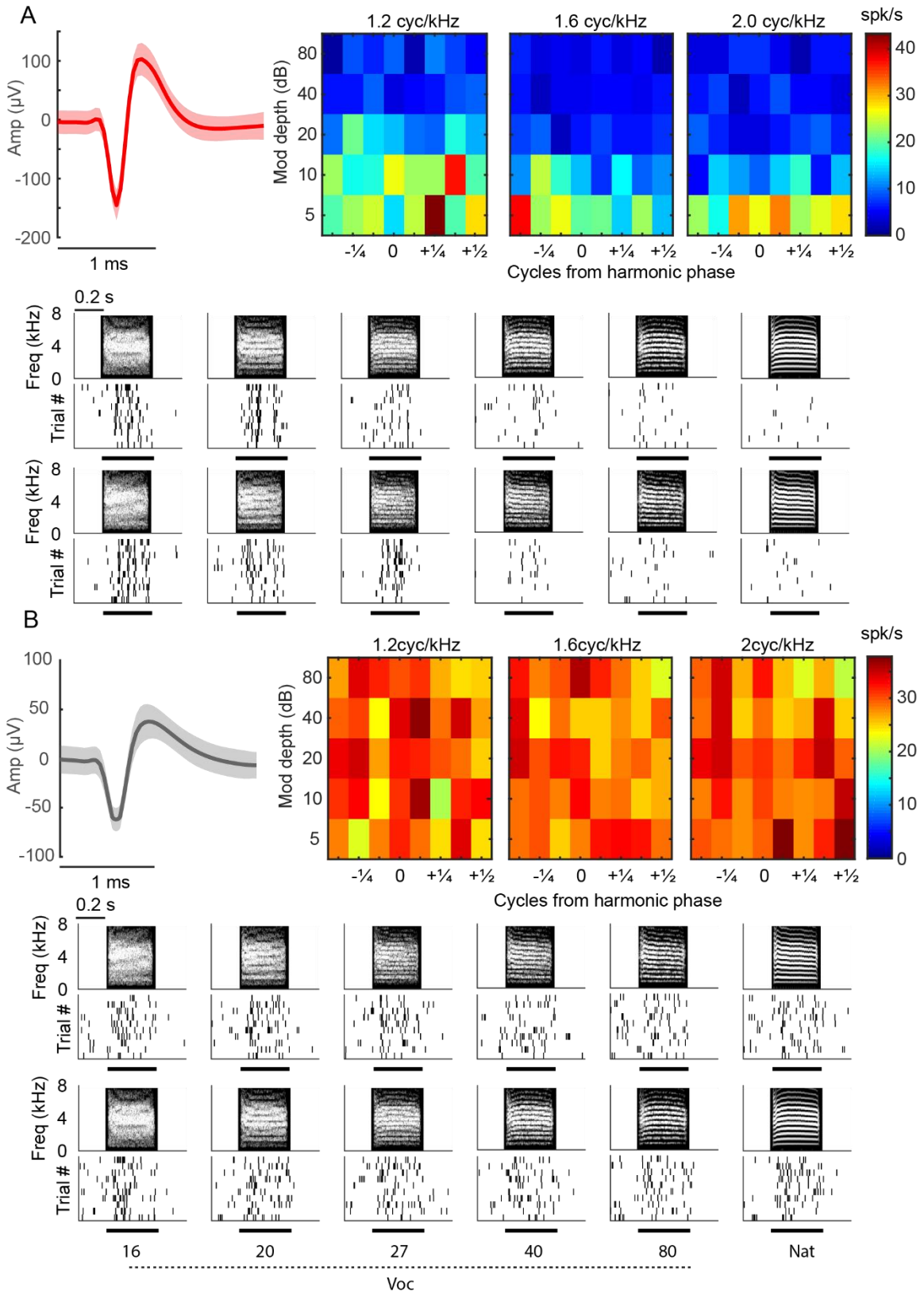


Figure 4.7 Call and ripple responses of deep region neurons with varying spectral selectivity.

** Figure and legend continued on next page*

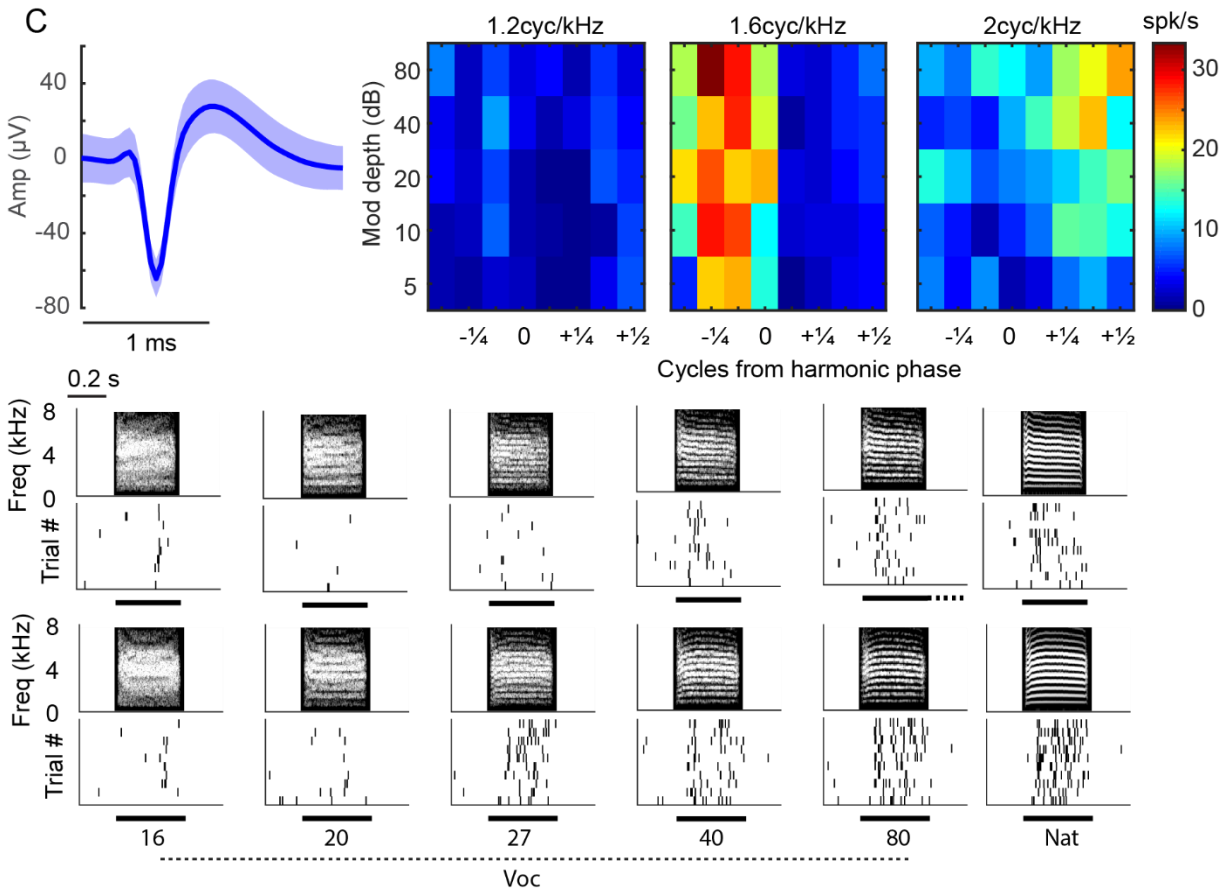


Figure 4.7 (continued). Responses of deep region neurons with varying spectral selectivity to ripples and calls. *(A)* Responses from an example low-resolution-selective (LS) neuron. Top-left plot shows the action potential waveform (mean ± SD) of the example neuron. Depth-phase matrices, which are heatmaps of the neuron's firing rates to individual ripples, are shown to the right of the action potential waveform. Raster plots (bottom) show the timing of spiking events before, during, and after the presentation of vocoded and natural calls. Two out of nine sets of vocoded (16 - 80 channels) and natural calls are shown as examples. Spectrograms of stimuli are shown above the raster plots. Black lines below each raster plot denote the duration of stimulus presentation. *(B)* Responses from an example unselective (US) neuron. Data are organized identically as in *(A)*. *(C)* Responses from an example high-resolution-selective (HS) neuron. Data are organized identically as in *(A)*.

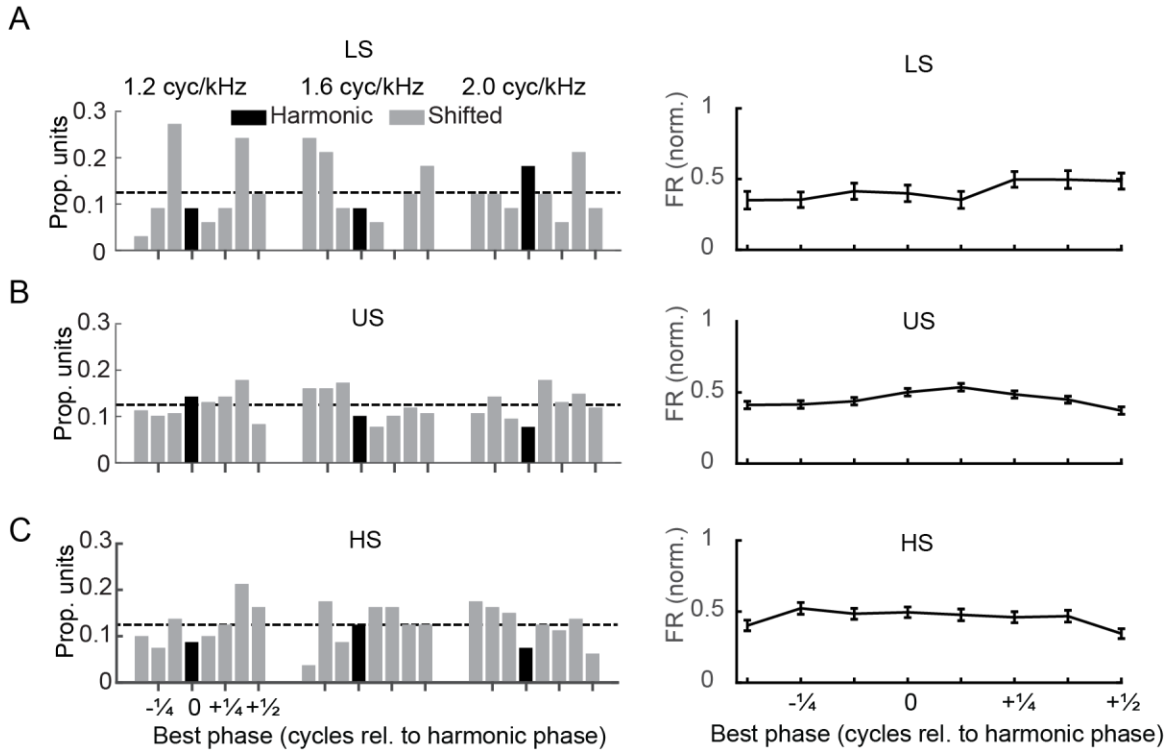


Figure 4.8 Deep region neurons do not show an enhanced representation of harmonic ripples. *(A, B, C) Left:* Distribution of best phase (the phase that elicited maximal response averaged across depths) for the three tested modulation densities. Distributions are shown separately for *(A)* low-resolution-selective (LS), *(B)* unselective (US), and ***(****C)* high-resolution-selective (HS) neurons. *Right:* Average phase response curves for *(A)* LS, *(B)* NS, and *(C)* HS neurons. Phase response curves were taken at each neurons' best modulation density and best depth, then scaled to range from 0 to 1.

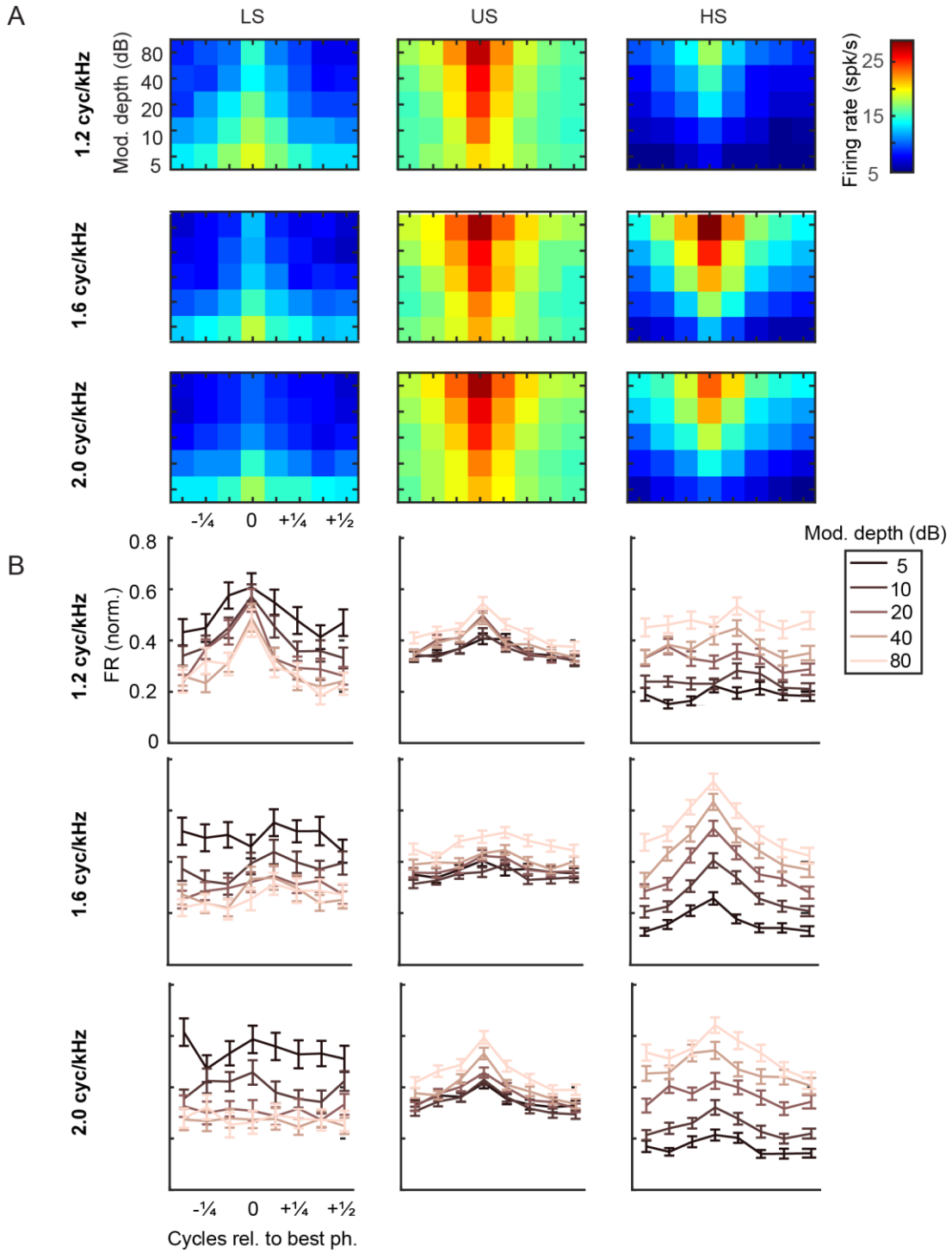


Figure 4.9 Deep region neurons' responses to ripples with varying density, depth, and phase. *(A)* Depth-phase matrices for low-resolution-selective (LS, left), unselective (US, middle), and high-resolution-selective (HS, right) neurons for ripples with modulation densities of 1.2 (top), 1.6 (middle), and 2.0 cyc/kHz (bottom). *(B)* Average normalized response curves (mean ± SE) showing responses of LS, US, and HS neurons to ripples. Data organized as in *(A)*. Line color indicates ripple modulation depth. Depth-phase matrices and response curves were centered at each neurons' best phase (best phase = zero).

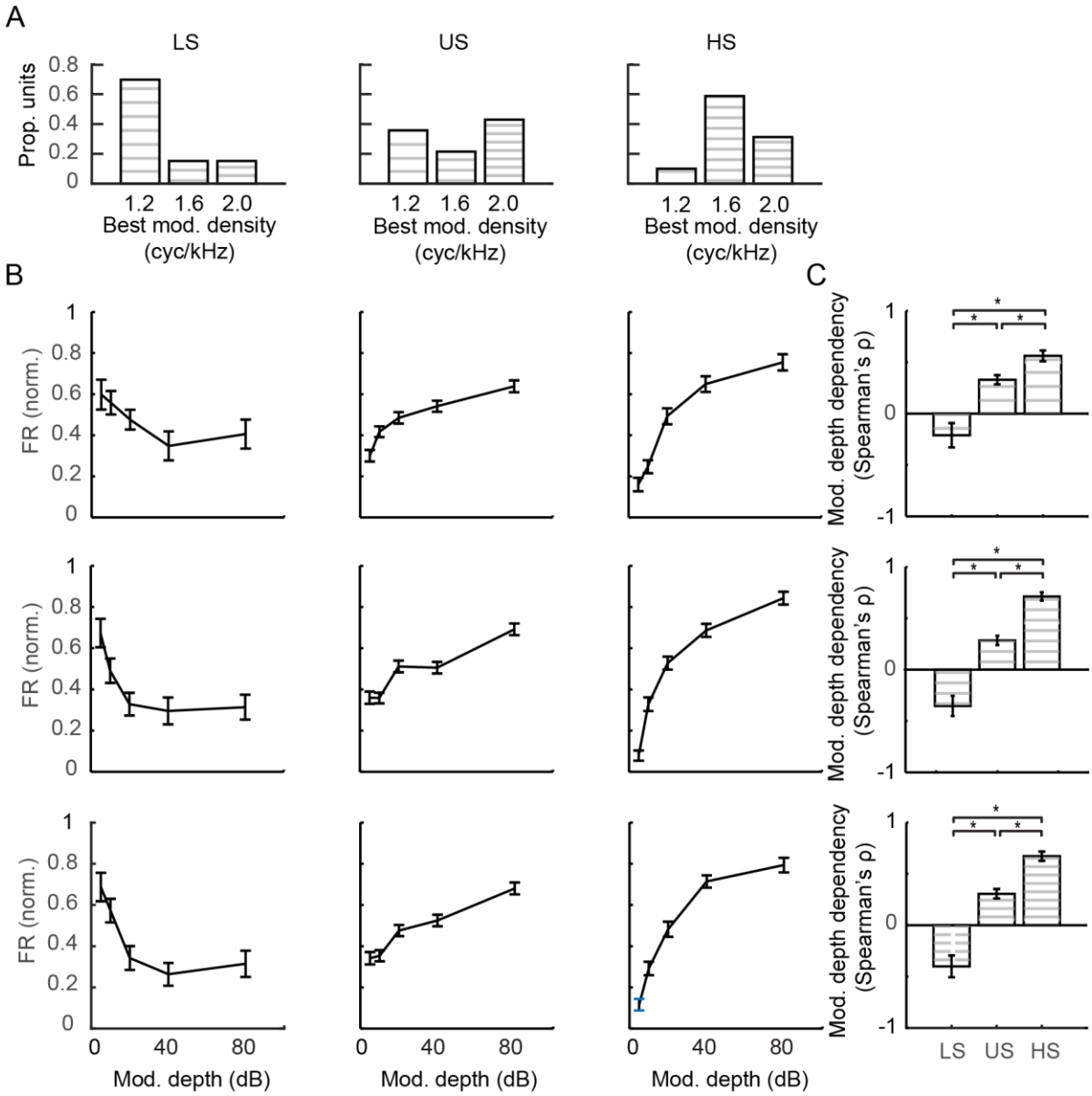


Figure 4.10 Neurons with differing spectral selectivity for calls are differentially modulated by ripple density and depth.

(A) Proportions of LS, US and HS neurons showing maximal response to each modulation density tested. *(B)* Normalized firing rates (mean ± SEM) of LS, US, and HS neurons evoked by each modulation depth, measured at modulation densities of 1.2, 1.6, and 2.0 cyc/kHz and at each neuron's best phase. Firing rates were normalized to range from 0 to 1 for each unit. *(C)* Modulation depth dependencies (mean ± SEM) of LS, US, and HS units, measured at each modulation density and at each neuron's best phase. Positive ρ indicates that driven responses increase with modulation depth, and negative ρ indicates that driven responses decrease with increases in modulation depth. * $p < 0.001$, ANOVA followed by Tukey's tests

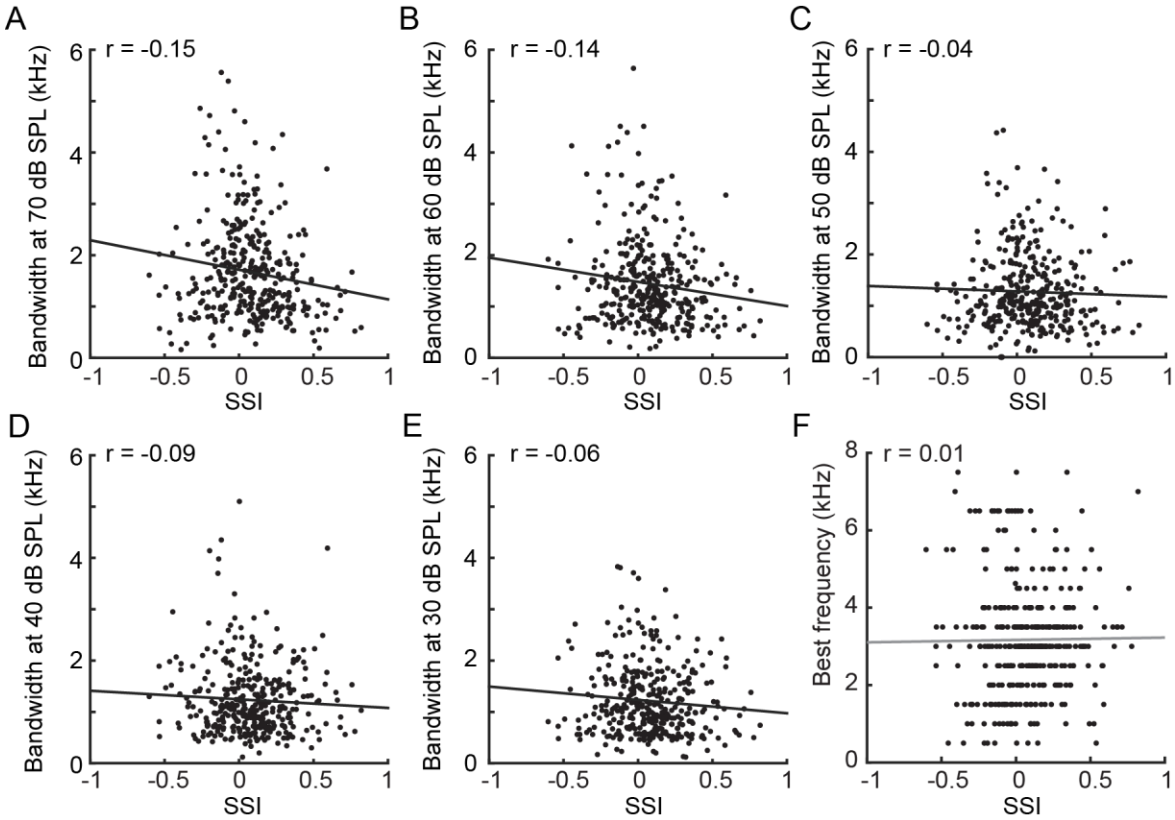


Figure 4.11 Correlation between call spectral selectivity index (SSI) and tone response properties.

(A-E) Scatter plots showing the relationship between SSI and bandwidths (widths of frequency response curve at half of the maximal height) measured at *(A)* 70 dB SPL, *(B)* 60 dB SPL, *(C)* 50 dB SPL, *(D)* 40 dB SPL, and *(E)* 30 dB SPL. Pearson's r correlation coefficients are shown on the top left of each plot. Significant negative correlations were observed between SSI and bandwidths at 60 dB SPL and 70 dB SPL ($p < 0.01$). *(F)* Scatter plot showing the relationship between SSI and best frequency measured from tones. Pearson's r correlation coefficients are shown on the top left of scatter plots.

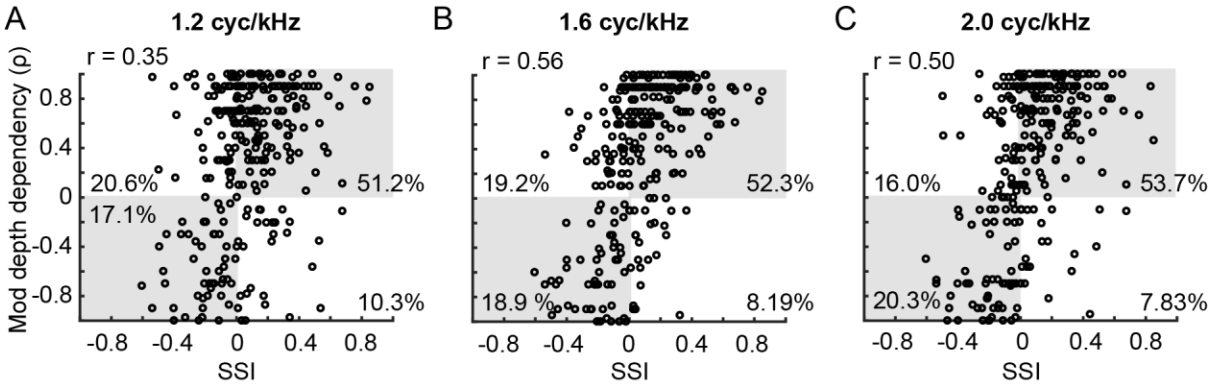


Figure 4.12 Correlation between spectral selectivity index (SSI) for calls and sensitivity to modulation depth of ripples.

(A-C) Scatter plots showing the relationship between SSI and modulation depth of ripples with modulation densities of *(A)* 1.2 cyc/kHz, *(B)* 1.6 cyc/kHz, and *(C)* 2.0 cyc/kHz. Shaded quadrants include units whose direction of modulation depth sensitivity matched the sign of call SSI; the percentage of units that fall in these quadrants are indicated. Pearson's r correlation coefficient is shown on the top left of each scatter plot. Correlation between SSI and modulation depth dependency was significant at all three modulation densities (all $p < 0.001$; N = 281 ripple- and call-responsive units).

4.3.4 Temporal response properties vary with spectral structure selectivity

To provide insight on the mechanisms of high-resolution selectivity in deep-region neurons, we identified temporal response properties that co-varied with SSI. We compared HS, US and LS units in three measures of response dynamics: 1) latencies to first spike during call presentation; 2) population peri-stimulus time histograms (pPSTHs) of responses to call stimuli; and 3) onset index of call responses.

First spike latencies of responses to calls differed significantly between groups while controlling for onset firing rates (**Figure 4.13C,** ANCOVA, $F(2, 404) = 6.46$, $p < 0.01$). Call response latencies were significantly longer in HS and NS units than in US units (Tukey's tests, $p < 0.05$). A similar pattern of latency differences can be seen in responses to pure tones (**Figure 4.13A**) and to ripples (**Figure 4.13B**), but these differences were not statistically significant after controlling for onset responses (ANCOVA, $p > 0.05$). Response latencies to calls were highly correlated with latencies to pure tones (**Figure 4.13D;** Pearson's $r = 0.59$, $p < 0.001$) and spectrally modulated ripples (**Figure 4.13E;** Pearson's $r = 0.75$, $p < 0.001$). Because calls, tones, and ripples have divergent acoustic features, the observation that latencies to these stimuli are highly intercorrelated suggests that latency differences are not due to neurons' preferences for particular time-varying acoustic features in calls.

The pPSTHs of responses to natural and vocoded calls differed between HS, US and LS units (e.g. **Figure 4.14B**). The population of US neurons showed a strong response that was aligned with stimulus onset, followed by a weaker, sustained response thereafter. In contrast, the population responses of HS and LS neurons were not the strongest at stimulus onset. We computed an onset index that weighed the onset response against sustained (post-onset) responses (Schumacher et al., 2011; X. Wang et al., 2005). We defined the onset period as the first 50 ms

following stimulus onset, and the sustained period as the following period until stimulus offset (**Figure 4.14A**). Onset index ranged from -1 to 1, with more positive values indicating stronger onset responses than sustained responses, and negative values indicating stronger sustained responses than onset responses. Onset index values were significantly higher in US neurons than in HS or LS neurons (**Figure 4.14C**, ANOVA, $F(2, 408) = 19.15$, $P < 0.001$; Tukey's tests, $P < 0.05$). These results showed that temporal response patterns were correlated with spectral selectivity; neurons that were insensitive to spectral resolution (US) had onset-dominant responses, while those that were sensitive to spectral resolution (LS and HS) exhibited stronger sustained responses.

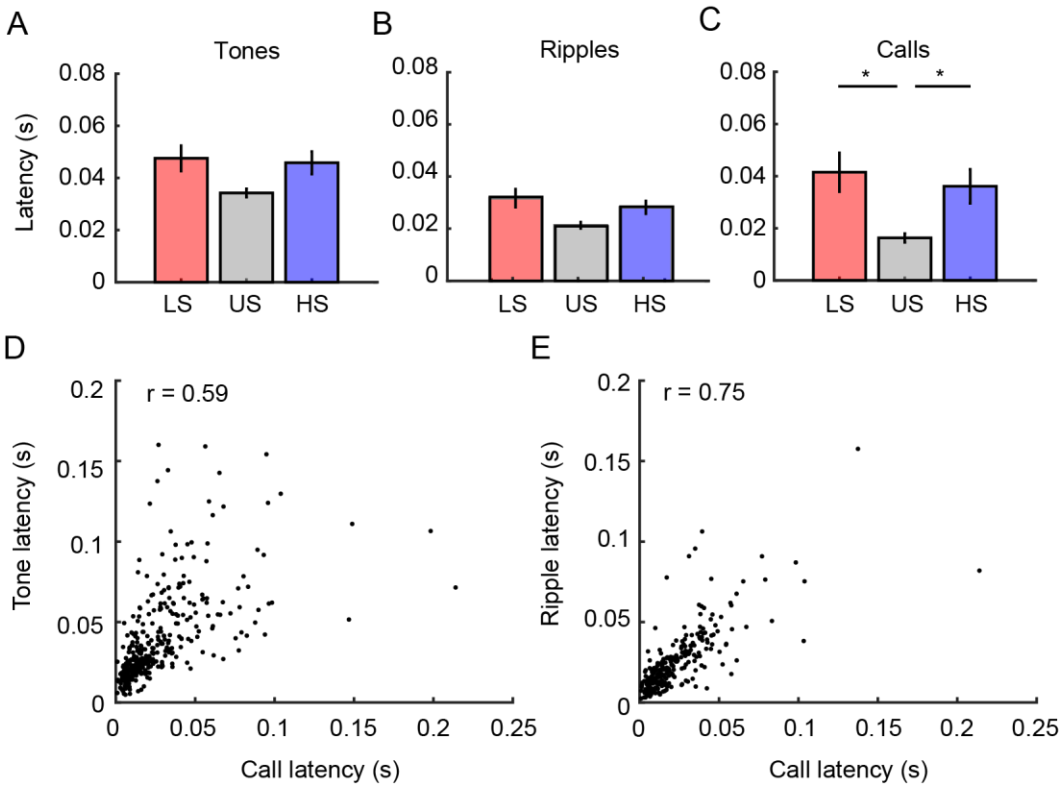


Figure 4.13 Response latencies vary with spectral selectivity and are consistent across stimulus types.
(A-C) First-spike latencies in response to *(A)* pure tones, *(B)* spectrally modulated ripples, and *(C)* natural and vocoded calls. Latencies are shown for NS, US and HS neurons (one-way ANCOVA with Tukey tests, * $p < 0.05$). *(D)* Scatter plot showing the relationship between call response latencies and tone response latencies (N = 355 tone- and call-responsive neurons). Pearson's r correlation coefficient is shown on the top left of scatter plot. *(E)* Scatter plots showing the relationship between call response latencies and ripple response latencies (N = 281 ripple- and call-responsive neurons). Pearson's r correlation coefficient is shown on the top left of scatter plot.

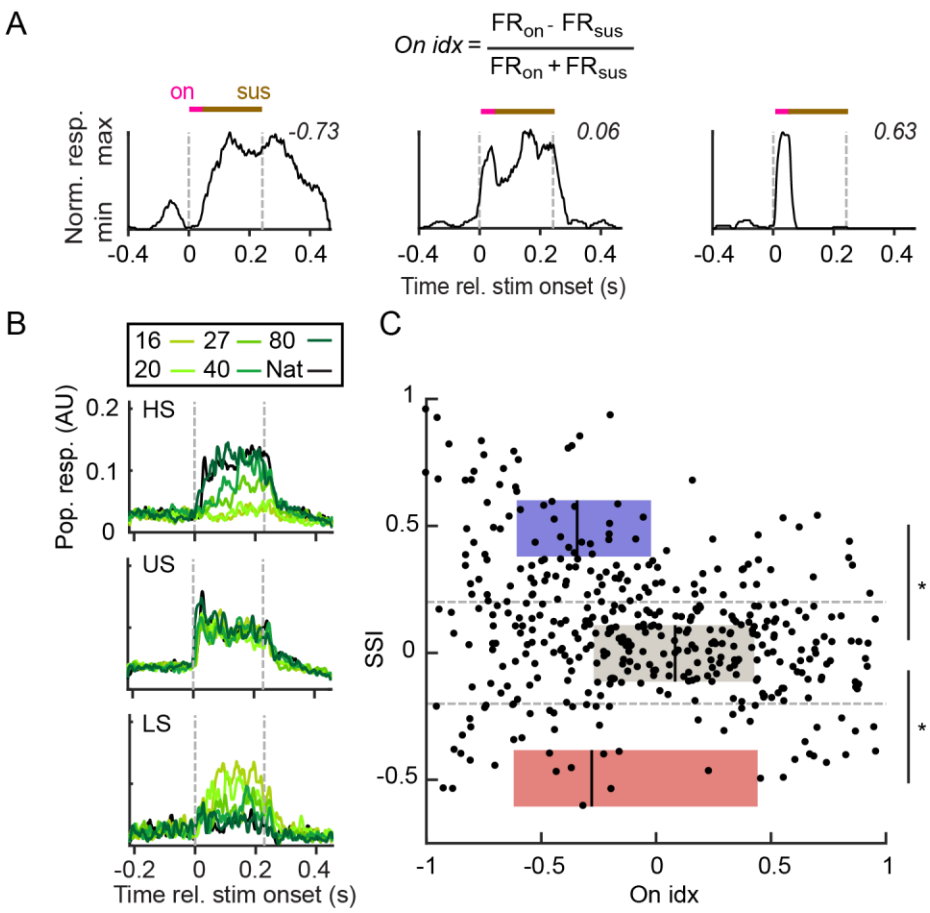


Figure 4.14 Temporal dynamics of call responses differ by spectral selectivity.
(A) Quantification of onset index and example PSTHs of single units with negative (left), near-zero (middle), and positive (right) onset indices. The formula for calculating onset index is shown above the PSTHs. Colored bars above each PSTH indicate the onset (pink) and sustained (brown) periods used to calculate onset index. *(B)* Example call-evoked pPSTHs for HS, US and NS neurons. pPSTHs to different stimulus groups are color-coded according to the legend. Responses of US neurons peak at stimulus onset, while HS and NS responses increase after stimulus-onset and are sustained over the duration of the stimulus. *(C)* Scatter plot and box plots showing the relationship between onset index and SSI. Onset responses are stronger in US neurons than in HS or LS neurons (one-way ANOVA with Tukey tests, * $p < 0.05$).

4.4 DISCUSSION

High-spectral-resolution selective neurons in the deep region maintained their selectivity in the context of song, showing population responses that closely tracked spectral parameters of song. Therefore, neural sensitivity to spectral resolution of vocoded calls generalizes to the processing of natural variations in spectral structure of vocalizations. Further, spectral selectivity was maintained across acoustic contexts – it is observed when vocal elements are presented in a dynamic sequence, as well as when they are presented in isolation. HS responses were lower than US responses during non-preferred song segments, and similar to US responses during preferred song segments.

Our analysis of neural responses to spectrally modulated ripples showed that spectrally selective neurons preferred sounds with deeper and denser modulations, but showed no preference for harmonic placement of spectral peaks. These neural response properties are consistent with our findings on birds' behavioral responsiveness in **Chapter 3** – birds responded similarly to harmonic and inharmonic calls, indicating that the preservation of deep spectral modulations was sufficient to elicit vocal responses.

Previous studies, reviewed further in Section 1.5.2, have characterized AC neurons by their responses to spectral modulation density, depth, and phase (Schreiner & Calhoun, 1994; Shamma et al., 1995). Other studies have reported anatomical organization of complex response properties, such as preferred bandwidth (Rauschecker et al., 1995), spectro-temporal modulation tuning (Hullett et al., 2016), and F0-specific responses (Bendor & Wang, 2005). Our study builds on previous work by providing evidence for robust anatomical grouping of neurons tuned to high modulation depth, and further establishing that modulation depth sensitivity contributes to the selective representation of behaviorally-relevant sounds. Though the deep region resides within a

primary sensory region and hence is unlikely to directly control behavioral output, neurons selective for modulation depth may send information to downstream regions for further processing, shaping behavioral sensitivity to high modulation depth signals.

To establish whether neurons in the deep region are involved in the modulation of vocal behavior by spectral resolution, it would be informative to inhibit neural activity in the deep region using pharmacological or electrolytic methods while birds are tested with the call-and-response behavioral paradigm. While the role of the deep region or other cortical regions in call-and-response behavior has not been previously tested, one study has investigated the role of a vocal control nucleus, the robustus archistriatalis (RA), in male zebra finches' vocal responses to distance calls. While male zebra finches normally respond more to female distance calls compared to male distance calls, males with RA lesions did not discriminate between female and male distance calls (Vicario et al., 2000). Interestingly, previous studies have identified deep region projections to areas that lie in close proximity to major nuclei of the vocal production pathway, including RA (Kelley & Nottebohm, 1979; Mello et al., 1998; Vates et al., 1996).

Our results can also be discussed with regard to the concept of contrast gain control and existing theories of natural sound processing. A previous study showed that AC neurons can dynamically adjust gain to compensate for changes in spectral contrast (variation in sound pressure across frequencies), but this compensation is incomplete (Rabinowitz et al., 2011). In our dataset, deep region responses clearly scaled with spectral modulation depth, indicating that if any compensatory gain control was present, it did not result in invariance to modulation depth. We characterized ripple responses at three spectral modulation densities, and found that HS neurons tend to prefer higher densities (1.6 cyc/kHz to 2.0 cyc/kHz). The preferred range corresponds roughly to modulations resulting from typical F0s of zebra finch calls, which range from ~500Hz

to ~700Hz (Mouterde et al., 2014; Vicario et al., 2001). While a more comprehensive sampling of the modulation space is needed to reach a firm conclusion, current results are consistent with the theory that the auditory system is sensitive to the statistical properties of vocalizations (Singh & Theunissen, 2003).

Given that deep region neurons did not prefer harmonic ripples (with clear F0) over inharmonic ripples (with ambiguous F0), we do not have evidence that they encode F0 by a firing rate code. Songbirds can perceive F0 from harmonic sounds with missing fundamentals (Cynx & Shapiro, 1986), indicating that F0 must be represented in the brain. It is possible that F0 is encoded outside of the auditory cortex or by auditory cortical neurons that we did not find because we did not sample enough neurons or cover its entire extent. F0 could also be encoded temporally by neurons' synchrony to F0-related envelope cues, as reported for avian midbrain neurons (Henry et al., 2017). However, cortical temporal representation of F0 is unlikely, given that these neurons can only synchronize to modulations up to 320 Hz (Knipschild et al., 1992), which is insufficient to encode the F0 of zebra finch calls (~500 Hz or above). It is possible that the temporal encoding of pitch occurs in subcortical structures, with that information being transformed into a rate code carried by AC neurons that remain to be identified.

Longer call response latencies in HS and LS neurons compared to US neurons suggest that input projections to these neurons may differ. In the songbird AC, deep-region neurons receive input from multiple pathways: 1) the superficial region (Vates et al., 1996); 2) the thalamorecipient regions (Y. Wang et al., 2010); and 3) the shell region of the auditory thalamus, a relatively sparse projection (Vates et al., 1996). Our results are consistent with a model of connectivity in which HS and LS neurons receive input from the superficial region, undergoing more intracortical processing than US neurons. The US neurons, in contrast, may receive input from the

thalamorecipient regions, and serve to encode the onsets of sounds regardless of spectral details. Sustained firing in the AC has been proposed to occur as a result of recurrent or interlaminar, intracortical processing, as well as feedback connections from higher cortical areas (X. Wang et al., 2005). Because HS and LS neurons show a more sustained firing profile than US neurons, intracortical processing could contribute to their selectivity. Consistent with this hypothesis, reciprocal connections between deep and superficial regions have been identified (Vates et al., 1996).

Chapter 5

CONCLUDING REMARKS

In this dissertation, I identified a behaviorally salient acoustic feature of vocal sounds and showed that it drives communication behavior and neural selectivity for vocalizations. Harmonic structure and deep spectral modulations both characterize call vocalizations in songbirds and the vocalizations of many other species. My studies established that deep spectral modulations are critical for behavioral responses to vocalizations, and are represented by a population of neurons in the auditory cortex as an attribute distinct from harmonic structure.

In Chapter 3, I showed that reduction of spectral resolution of communication calls, which diminishes both harmonic structure and spectral modulations, decreases the vocal responses evoked by these calls. While birds were sensitive to spectral degradation of calls, their responses were not impaired by disruption of harmonicity, indicating that vocal responses require the presence of distinct peaks and troughs in the power spectrum, but do not require frequency components to be harmonically related. I then reported the results of neurophysiological experiments where I identified an anatomically localized population of neurons in the auditory cortex whose responses decreased with spectral degradation of calls in a manner mirroring that of behavior. Spectral selectivity for behaviorally relevant calls was found in the deep output region of the primary auditory cortex, but not in the upstream thalamorecipient region. This finding suggested that spectral selectivity for behaviorally relevant calls is generated by neural mechanisms operating within the auditory cortex.

In Chapter 4, spectrally-selective neurons (those that show selectivity for calls with high spectral resolution) in the deep region were further found to maintain their selectivity when

processing complex, acoustically diverse song. In addition, deep region spectrally-selective neurons were sensitive to modulation depth and not to harmonicity. These response properties point to a connection between deep region spectral processing and birds' behavioral responses to vocalizations.

Our behavioral results indicating that birds' vocal responses are not impaired by inharmonicity was surprising, given that they are known to discriminate very small deviations from harmonicity (Lohr & Dooling, 1998). It is possible that harmonicity is not important in this behavioral context but is important for other listening situations. Further insights on how birds react to harmonic and inharmonic signals can be derived from an auditory preference test, where birds' tendency to actively elicit playbacks of different stimuli are compared. This experiment would establish whether birds show a preference for listening to harmonic calls over inharmonic calls, akin to the way that human listeners prefer to listen to consonant musical chords, whose spectra are more similar to harmonic structures, over dissonant musical chords (Bowling et al., 2018).

Two lines of inquiry on the neural processing of vocalizations arise from our studies, and can be further addressed by future studies in the songbird and in other vocal communicators. First, what are the neural mechanisms leading to sensitivity to modulation depth? Our studies of song responses indicate that while spectrally selective neurons in the deep region respond with similar magnitude as unselective neurons to preferred stimuli, their responses are suppressed relative to unselective neurons when processing non-preferred stimuli. This suggests that spectral selectivity could be shaped by inhibition that is recruited by non-preferred stimuli. A combination of electrophysiology and pharmacological techniques to locally manipulate inhibitory circuits could serve to test this hypothesis. Second, do spectrally selective neurons in the deep region have

distinct morphological properties and downstream projections? To answer this question, single-unit electrophysiology can be combined with injection of neural tracers into single cells following recording. Examining the somatic and dendritic morphologies of filled single units will elucidate whether spectral selectivity is associated with cell-specific morphological properties. By contrasting the anatomical projection targets of filled spectrally-selective and unselective neurons, one could gain further understanding on how spectrally-selective neurons could drive behavior by sending information to other processing centers in the brain.

BIBLIOGRAPHY

Alonso, R. G., Kopuchian, C., Amador, A., Suarez, M. D. L. A., Pablo, L., & Mindlin, G. B. (2016). Difference between the vocalizations of two sister species of pigeons explained in dynamical terms. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*, *202*(5), 361–370. https://doi.org/10.1007/s00359-016-1082-3.Difference

Amagai, S., Dooling, R. J., Kidd, T. L., & Lohr, B. (1999). Detection of modulation in spectral envelopes and linear-rippled noises by budgerigars (Melopsittacus undulatus). *Journal of the Acoustical Society of America*, *105*(3), 2029–2035.

Araki, M., Bandi, M. M., & Yazaki-Sugiyama, Y. (2016). Mind the gap: Neural coding of species identity in birdsong prosody. *Science*, *354*(6317), 1282–1287. https://doi.org/10.1126/science.aah6799

Attias, H., & Schreiner, C. E. (1997). Coding of Naturalistic Stimuli by Auditory Midbrain Neurons. *Advances in Neural Information Processing Systems*, *10*, 103–109.

Barbour, D. L., & Wang, X. (2003). Contrast tuning in auditory cortex. *Science*, *299*(5609), 1073–1075. https://doi.org/10.1126/science.1080425

Belin, P., Fecteau, S., & Be, C. (2004). Thinking the voice : neural correlates of voice perception, *8*(3). https://doi.org/10.1016/j.tics.2004.01.008

Belyk, M., & Brown, S. (2017). The origins of the vocal brain in humans. *Neuroscience and Biobehavioral Reviews*, *77*, 177–193. https://doi.org/10.1016/j.neubiorev.2017.03.014

Bendor, D., & Wang, X. (2005). The Neuronal Representation of Pitch in Primate Auditory Cortex. *Nature*, *436*(7054), 1161–1165.

Bowling, D. L., Purves, D., & Gill, K. Z. (2018). Vocal similarity predicts the relative attraction of musical chords, *115*(1). https://doi.org/10.1073/pnas.1713206115

Brainard, M. S. M., & Doupe, A. J. A. (2013). Translating Birdsong: Songbirds as a Model for Basic and Applied Medical Research. *Annual Review of Neuroscience*, *36*(1), 489–517. https://doi.org/10.1146/annurev-neuro-060909-152826

Bregman, M. R., Patel, A. D., & Gentner, T. Q. (2016). Songbirds use spectral shape, not pitch, for sound pattern recognition. *Proceedings of the National Academy of Sciences*, *113*(6), 1666–1671. https://doi.org/10.1073/pnas.1515380113

Burgering, M., Vroomen, J., & Ten Cate, C. (2018). Zebra Finches (Taeniopygia guttata) Can Categorize Vowel-Like Sounds on Both the Fundamental Frequency (“Pitch”) and Spectral Envelope, (September). https://doi.org/10.1037/com0000143

Calabrese, A., & Woolley, S. M. N. (2015). Coding principles of the canonical cortical

microcircuit in the avian brain. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), 3517–3522. https://doi.org/10.1073/pnas.1408545112

Catz, N., & Noreña, A. J. (2013). Enhanced representation of spectral contrasts in the primary auditory cortex. *Frontiers in Systems Neuroscience*, *7*(June), 1–16. https://doi.org/10.3389/fnsys.2013.00021

Cazau, D., Adam, O., Aubin, T., Laitman, J. T., & Reidenberg, J. S. (2016). A study of vocal nonlinearities in humpback whale songs : from production mechanisms to acoustic analysis. *Nature Publishing Group*, (June), 1–12. https://doi.org/10.1038/srep31660

Charrier, I., & Sturdy, C. B. (2005). Call-based species recognition in black-capped chickadees. *Behavioural Processes*, *70*, 271–281. https://doi.org/10.1016/j.beproc.2005.07.007

Chase, S. M., & Young, E. D. (2007). First-spike latency information in single neurons increases when referenced to population onset. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(12), 5175–5180. https://doi.org/10.1073/pnas.0610368104

Cynx, J., & Shapiro, M. (1986). Perception of Missing Fundamental by a Species of Songbird (Sturnus vulgaris). *Journal of Comparative Psychology*, *100*(4), 356–360.

Davies-venn, E., Nelson, P., & Souza, P. (2015). Comparing auditory filter bandwidths , spectral ripple modulation detection , spectral ripple discrimination , and speech recognition : Normal and impaired hearing a). *Journal of the Acoustical Society of America*, *138*(1), 493–503. https://doi.org/10.1121/1.4922700

Duifhuis, H., Willems, L. F., & Sluyter, R. J. (1982). Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *The Journal of the Acoustical Society of America*, *71*(6), 1568–1580. https://doi.org/10.1121/1.387811

Eddins, D. A., & Bero, E. M. (2007). bandwidth , and carrier frequency region Spectral modulation detection as a function of modulation frequency , carrier bandwidth , and carrier frequency region. *Journal of the Acoustical Society of America*, *121*(1), 363–372. https://doi.org/10.1121/1.2382347

Ehret, G., & Riecke, S. (2002). Mice and humans perceive multiharmonic communication sounds in the same way. *Proceedings of the National Academy of Sciences*, *99*(1), 479–482.

Elemans, C. P. H. (2014). The singer and the song : The neuromechanics of avian sound production. *Current Opinion in Neurobiology*, *28*, 172–178. https://doi.org/10.1016/j.conb.2014.07.022

Eliades, S. J., & Miller, C. T. (2016). Marmoset Vocal Communication : Behavior and Neurobiology, 286–299. https://doi.org/10.1002/dneu.22464

Elie, J. E., & Theunissen, F. E. (2016). The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. *Animal Cognition*, *19*(2), 285–315. https://doi.org/10.1007/s10071-015-0933-6

Elliott, T. M., & Theunissen, F. E. (2009). The Modulation Transfer Function for Speech Intelligibility. *PLoS Computational Biology*, *5*(3), 1–14. https://doi.org/10.1371/journal.pcbi.1000302

Fee, M. S., Shraiman, B., & Mitra, P. P. (1998). The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird, *395*(September), 67–71.

Feng, L., & Wang, X. (2017). Harmonic template neurons in primate auditory cortex underlying complex sound processing. *Proceedings of the National Academy of Sciences*, *114*(5), E840–E848. https://doi.org/10.1073/pnas.1607519114

Fernández-vargas, M., & Johnston, R. E. (2015). Ultrasonic Vocalizations in Golden Hamsters (Mesocricetus auratus) Reveal Modest Sex Differences and Nonlinear Signals of Sexual Motivation, 1–29. https://doi.org/10.1371/journal.pone.0116789

Fitch, W. T., Neubauer, R., & Herzel, H. (2002). Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production, 407–418. https://doi.org/10.1006/anbe.2001.1912

Fu, Q., Chinchilla, S., Nogaki, G., & Iii, J. J. G. (2005). Voice gender identification by cochlear implant users : The role of spectral and temporal resolution Voice gender identification by cochlear implant users :, (October). https://doi.org/10.1121/1.1985024

Gaudrain, E. (2016). Vocoder, v1.0. https://doi.org/10.5281/zenodo.48120.

Gonzalez, J., & Oliver, J. C. (2005). Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *Journal of the Acoustical Society of America*, *118*(1), 461–470. https://doi.org/10.1121/1.1928892

Hall, M. L. (2004). A review of hypotheses for the functions of avian duetting. *Behav Ecol Sociobiol*, 415–430. https://doi.org/10.1007/s00265-003-0741-x

Harris, K. D., & Mrsic-Flogel, T. D. (2013). Cortical connectivity and sensory coding. *Nature*. https://doi.org/10.1038/nature12654

Hauber, M. E., Campbell, D. L. M., & Woolley, S. M. N. (2010). The functional role and female perception of male song in Zebra Finches. *Emu*, *110*(3), 209–218. https://doi.org/10.1071/MU10003

Heffner, H., & Whitfield, I. (1976). Perception of the missing fundamental by cats. *Journal of the Acoustical Society of America*, *59*(4), 915–919. https://doi.org/10.1121/1.380951

Henry, K. S., Abrams, K. S., Forst, J., Mender, M. J., Nelans, E. G., Idrobo, F., & Carney, L. H. (2017). Midbrain Synchrony to Envelope Structure Supports Behavioral Sensitivity to Single-Formant Vowel-Like Sounds in Noise. *Journal of the Association for Research in Otolaryngology: JARO*, *18*, 165–181. https://doi.org/10.1007/s10162-016-0594-4

Holveck, M. J., & Riebel, K. (2007). Preferred songs predict preferred males: consistency and

repeatability of zebra finch females across three test contexts. *Animal Behaviour*, *74*, 297–309. https://doi.org/10.1016/j.anbehav.2006.08.016

Houtsma, A. J. M., & Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *Journal of the Acou*, *87*(1), 304–310.

Hullett, P. W., Hamilton, X. L. S., Mesgarani, N., Schreiner, X. C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli, *36*(6), 2014–2026. https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Janik, V. M. (2014). Cetacean vocal learning and communication. *Current Opinion in Neurobiology*, *28*, 60–65. https://doi.org/10.1016/j.conb.2014.06.010

Joyce, H., Peter, L., & Angela, S. (2005). Elephants are capable of vocal learning. *Nature*, *434*, 455–456.

Kelley, D. B. (2004). Vocal communication in frogs. *Current Opinion in Neurobiology*, *14*, 751–757. https://doi.org/10.1016/j.conb.2004.10.015

Kelley, D. B., & Nottebohm, F. (1979). Projections of a Telencephalic Auditory Nucleus- Field L- in the Canary. *Journal of Comparative Neurology*, *183*, 455–470.

Kikuchi, Y., Horwitz, B., Mishkin, M., & Rauschecker, J. P. (2014). Processing of harmonics in the lateral belt of macaque auditory cortex. *Frontiers in Neuroscience*, *8*(July), 204. https://doi.org/10.3389/fnins.2014.00204

Knipschild, M., Dorrscheidt, G. J., & Rubsamen, R. (1992). Setting complex tasks to single units in the avian auditory forebrain . I: Processing of complex artificial stimuli. *Hearing Research*, *57*, 216–230.

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., & Brefczynski-Lewis, J. A. (2009). Human Cortical Organization for Processing Vocalizations Indicates Representation of Harmonic Structure as a Signal Attribute. *Journal of Neuroscience*, *29*(7), 2283–2296. https://doi.org/10.1523/JNEUROSCI.4145-08.2009

Lim, Y., Lagoy, R., Shinn-Cunningham, B. G., & Gardner, T. J. (2016). Transformation of temporal sequences in the zebra finch auditory system. *ELife*, *5*(NOVEMBER2016). https://doi.org/10.7554/eLife.18205

Lohr, B., & Dooling, R. J. (1998). Detection of Changes in Timbre and Harmonicity in Complex Sounds by Zebra Finches (Taeniopygia guttata) and Budgerigars. *Journal of Comparative Psychology*, *112*(1), 36–47.

Maat, A. Ter, Trost, L., Sagunsky, H., Seltmann, S., & Gahr, M. (2014). Zebra Finch Mates Use Their Forebrain Song System in Unlearned Call Communication. *PloS One*, *9*(10). https://doi.org/10.1371/journal.pone.0109334

Marler, P. (2004). Bird calls : a cornucopia for communication. In P. Marler & H. Slabbekoorn (Eds.), *Nature's Music: The Science of Birdsong* (pp. 133–177).

McDermott, J. H., Ellis, D. P. W., & Kawahara, H. (2012). Inharmonic Speech: A Tool for the Study of Speech Perception and Separation. *Proc. SAPA-SCALE 2012*, 114–117. Retrieved from http://www.ee.columbia.edu/~dpwe/pubs/McDEK12-inharmonic.pdf

McPherson, M. J., & McDermott, J. H. (2018). Diversity in pitch perception revealed by task dependence. *Nature Human Behavior*, *2*(1), 52–66. https://doi.org/10.1038/s41562-017-0261-8.Diversity

Meliza, C. D., & Margoliash, D. (2012). Emergence of selectivity and tolerance in the avian auditory cortex. *Journal of Neuroscience*, *32*(43), 15158–15168. https://doi.org/10.1523/JNEUROSCI.0845-12.2012

Mello, C. V., Vates, G. E., Okuhata, S., & Nottebohm, F. (1998). Descending auditory pathways in the adult male zebra finch (Taeniopygia guttata). *Journal of Comparative Neurology*, *395*(2), 137–160. https://doi.org/10.1002/(SICI)1096-9861(19980601)395:2<137::AID-CNE1>3.0.CO;2-3

Mormann, F., Kornblith, S., Quiroga, R. Q., Kraskov, A., Cerf, M., & Fried, I. (2008). Latency and Selectivity of Single Neurons Indicate Hierarchical Processing in the Human Medial Temporal Lobe, *28*(36), 8865–8872. https://doi.org/10.1523/JNEUROSCI.1640-08.2008

Mouterde, S. C., Theunissen, F. E., Elie, J. E., Vignal, C., & Mathevon, N. (2014). Acoustic Communication and Sound Degradation: How Do the Individual Signatures of Male and Female Zebra Finch Calls Transmit over Distance? *PloS One*, *9*(7), e102842. https://doi.org/10.1371/journal.pone.0102842

Noll, M. A. (1964). Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection. *The Journal of the Acoustical Society of America*, *36*, 296. https://doi.org/https://doi.org/10.1121/1.1918949

Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(50), 19451–19469. https://doi.org/10.1523/JNEUROSCI.2880-13.2013

Norman-haignere, X. S. V, Albouy, P., Caclin, A., Mcdermott, J. H., Kanwisher, N. G., & Tillmann, B. (2016). Pitch-Responsive Cortical Regions in Congenital Amusia, *36*(10), 2986–2994. https://doi.org/10.1523/JNEUROSCI.2705-15.2016

O'Connor, K. N., Barruel, P., & Sutter, M. L. (2000). Global processing of spectrally complex sounds in macaques (Macaca mullata) and humans. *J Comp Physiol A*, *186*, 903–912. https://doi.org/10.1007/s003590000145

Osmanski, M. S., Marvit, P., Depireux, D., & Dooling, R. (2009). Discrimination of auditory gratings in birds. *Hearing Research*, *256*(0), 11–20.

https://doi.org/10.1016/j.heares.2009.04.020.Discrimination

Owren, M. J. (2002). Nonlinear analysis of irregular animal vocalizations, *111*(6). https://doi.org/10.1121/1.1474440

Penagos, H., Melcher, J. R., & Oxenham, A. J. (2004). A Neural Representation of Pitch Salience in Nonprimary Human Auditory Cortex Revealed with Functional Magnetic Resonance Imaging. *The Journal of Neuroscience*, *24*(30), 6810–6815. https://doi.org/10.1523/JNEUROSCI.0383-04.2004

Perez, E. C., Elie, J. E., Boucaud, I. C. a., Crouchet, T., Soulage, C. O., Soula, H. a., … Vignal, C. (2015). Physiological resonance between mates through calls as possible evidence of empathic processes in songbirds. *Hormones and Behavior*, *75*, 130–141. https://doi.org/10.1016/j.yhbeh.2015.09.002

Phillips, D. P. (2000). Factors shaping the response latencies of neurons in the cat ' s auditory cortex, *93*(1998), 33–41.

Plack, C., & Oxenham, A. (2005). The Psychophysics of Pitch. In C. Plack, A. Oxenham, R. Fay, & A. Popper (Eds.), *Pitch: Neural Coding and Perception* (pp. 7–55).

Popham, S., Boebinger, D., Ellis, D. P. W., Kawahara, H., & McDermott, J. H. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature Communications*, *9*(1), 2122. https://doi.org/10.1038/s41467-018-04551-8

Prat, Y., Taub, M., & Yovel, Y. (2015). Vocal learning in a social mammal : Demonstrated by isolation and playback experiments in bats. *Sci. Adv.*, 1–5.

Puschmann, S., Uppenkamp, S., Kollmeier, B., & Thiel, C. M. (2010). NeuroImage Dichotic pitch activates pitch processing centre in Heschl ' s gyrus. *NeuroImage*, *49*(2), 1641–1649. https://doi.org/10.1016/j.neuroimage.2009.09.045

Quiroga, R. Q., Nadasdy, Z., & Ben-Shaul, Y. (2004). Unsupervised Spike Detection and Sorting with Wavelets and Superparamagnetic Clustering. *Neural Computation*, *16*(8), 1661–1687. https://doi.org/10.1162/089976604774201631

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., & King, A. J. (2011). Contrast Gain Control in Auditory Cortex. *Neuron*, *70*(6), 1178–1191. https://doi.org/10.1016/j.neuron.2011.04.030

Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of Complex Sounds in the Macaque Nonprimary Auditory Cortex. *Science*, *268*(April).

Rendall, D., Rodman, P. S., & Emond, R. E. (1996). Vocal recognition of individuals and kin in free-ranging rhesus monkeys. *Animal Behavior*, *51*, 1007–1015.

Riede, T., & Goller, F. (2010). Peripheral mechanisms for vocal production in birds - differences and similarities to human speech and singing. *Brain and Language*, *115*(1), 69–80.

https://doi.org/10.1016/j.bandl.2009.11.003

Rieke, F., Bodnar, D., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society B: Biological Sciences*, *262*, 259–265.

Sabin, A. T., Eddine, D. A., & Wright, B. A. (2012). Perceptual learning of auditory spectral modulation detection. *Exp Brain Res.*, *218*(4), 567–577. https://doi.org/10.1007/s00221-012-3049-0.Perceptual

Scheffers, M. T. (1983). Simulation of auditory analysis of pitch: an elaboration on the DWS pitch meter. *The Journal of the Acoustical Society of America*, *74*(6), 1716–1725. https://doi.org/10.1121/1.390280

Schneider, D. M., & Woolley, S. M. N. (2013). Sparse and background-invariant coding of vocalizations in auditory scenes. *Neuron*, *79*(1), 141–152. https://doi.org/10.1016/j.neuron.2013.04.038

Schreiner, C. E., & Calhoun, B. M. (1994). Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions. *Auditory Neuroscience*, *1*, 39–61.

Schumacher, J. W., Schneider, D. M., & Woolley, S. M. N. (2011). Anesthetic state modulates excitability but not spectral tuning or neural discrimination in single auditory midbrain neurons. *Journal of Neurophysiology*, *106*(2), 500–514. https://doi.org/10.1152/jn.01072.2010

Seyfarth, R. M., & Cheney, D. L. (2017). The origin of meaning in animal signals. *Animal Behaviour*, *124*, 339–346. https://doi.org/10.1016/j.anbehav.2016.05.020

Shamma, S. A. (1985). Speech processing in the auditory system. II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, *78*(5), 1622–1632. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3840813

Shamma, S. A., Fleshman, J. W., Wiser, P. R., & Versnel, H. (1993). Organization of Response Areas in Ferret Primary Auditory Cortex. *Journal of Neurophysiology*, *69*(2), 368–383.

Shamma, S. A., & Klein, D. (2000). The case of the missing pitch templates : How harmonic templates emerge in the early auditory system. *Journal of the Acoustical Society of America*, *107*(5), 2631–2644. https://doi.org/10.1121/1.428649

Shamma, S. A., Versnel, H., & Kowalski, N. (1995). Ripple Analysis in Ferret Primary Auditory Cortex. I. Response Characteristics of Single Units to Sinusoidally Rippled Spectra. *Auditory Neuroscience*, *1*, 233–254.

Shannon, R. V, Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with primarily temporal cues. *Science*, *270*(5234), 303–304. https://doi.org/10.1126/science.1115233

Shipley, C., Carterette, E. C., Buchwald, J. S., Shipley, C., Carterette, E. C., & Buchwald, J. S. (2005). The effects of articulation on the acoustical structure of feline vocalizations The effects of feline of articulation on the acoustical structure vocalizations, *902*(1991). https://doi.org/10.1121/1.1894652

Simmons, J. A., & Simmons, A. M. (2011). Bats and frogs and animals in between: Evidence for a common central timing mechanism to extract periodicity pitch. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*. https://doi.org/10.1007/s00359-010-0607-4

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394. https://doi.org/10.1121/1.1624067

Soltis, J. (2010). Vocal communication in African elephants (Loxodonta africana). *Zoo Biology*, *29*(2), 192–209. https://doi.org/10.1002/zoo.20251

Song, X., Osmanski, M. S., Guo, Y., & Wang, X. (2016). Complex pitch perception mechanisms are shared by humans and a New World monkey, *113*(3), 781–786. https://doi.org/10.1073/pnas.1516120113

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., & Mitra, P. P. (2000). A procedure for an automated measurement of song similarity. *Animal Behavior*, *59*, 1–12. https://doi.org/10.1006/anbe.1999.1416

Ter-mikaelian, M., Semple, M. N., & Sanes, D. H. (2018). Effects of spectral and temporal disruption on cortical encoding of gerbil vocalizations, 1190–1204. https://doi.org/10.1152/jn.00645.2012

Terleph, T. a, Mello, C. V, & Vicario, D. S. (2006). Auditory topography and temporal response dynamics of canary caudal telencephalon. *Journal of Neurobiology*, *66*(3), 281–292. https://doi.org/10.1002/neu.20219

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews. Neuroscience*, *15*(6), 355–366. https://doi.org/10.1038/nrn3731

Thompson, W. F., & Balkwill, L. (2006). Decoding speech prosody in five languages. *Semiotica*, 407–424.

Titze, I. R. (2017). Human speech: a restricted use of the mammalian larynx, *31*(2), 135–141. https://doi.org/10.1016/j.jvoice.2016.06.003

Tsukano, H., Horie, M., Ohga, S., Takahashi, K., & Kubota, Y. (2017). Reconsidering Tonotopic Maps in the Auditory Cortex and Lemniscal Auditory Thalamus in Mice, *11*(February), 1–8. https://doi.org/10.3389/fncir.2017.00014

Vates, G. E., Broome, B. M., Mello, C. V., & Nottebohm, F. (1996). Auditory pathways of caudal telencephalon and their relation to the song system of adult male zebra finches (Taenopygia

guttata). *Journal of Comparative Neurology*, *366*, 613–642. https://doi.org/10.1002/(SICI)1096-9861(19960318)366:4<613::AID-CNE5>3.0.CO;2-7

Vicario, D. S., Naqvi, N. H., & Raksin, J. N. (2000). Behavioral Discrimination of Sexually Dimorphic Calls by Male Zebra Finches Requires an Intact Vocal Motor Pathway ABSTRACT :, 109–120.

Vicario, D. S., Naqvi, N. H., & Raksin, J. N. (2001). Sex differences in discrimination of vocal communication signals in a songbird. *Animal Behaviour*, *61*(4), 805–817. https://doi.org/10.1006/anbe.2000.1651

Vignal, C., & Mathevon, N. (2011). Effect of acoustic cue modifications on evoked vocal response to calls in zebra finches (Taeniopygia guttata). *Journal of Comparative Psychology (Washington, D.C. : 1983)*, *125*(2), 150–161. https://doi.org/10.1037/a0020865

Vignal, C., Mathevon, N., & Mottin, S. (2004). Audience drives male songbird response to partner's voice. *Nature*, *430*(6998), 448–451. https://doi.org/10.1038/nature02645

Vignal, C., Mathevon, N., & Mottin, S. (2008). Mate recognition by female zebra finch: analysis of individuality in male call and first investigations on female decoding process. *Behavioural Processes*, *77*(2), 191–198. https://doi.org/10.1016/j.beproc.2007.09.003

Walker, K. M. M., Bizley, J. K., King, A. J., & Schnupp, J. W. H. (2011). Cortical encoding of pitch: recent results and open questions. *Hearing Research*, *271*(1–2), 74–87. https://doi.org/10.1016/j.heares.2010.04.015

Wang, X. (2013). The harmonic organization of auditory cortex. *Frontiers in Systems Neuroscience*, *7*(December), 114. https://doi.org/10.3389/fnsys.2013.00114

Wang, X., Lu, T., Snider, R. K., & Liang, L. (2005). Sustained firing in auditory cortex evoked by preferred stimuli. *Nature*, *435*(7040), 341–346. https://doi.org/10.1038/nature03565

Wang, X., & Walker, K. M. M. (2012). Neural Mechanisms for the Abstraction and Use of Pitch Information in Auditory Cortex. *Journal of Neuroscience*, *32*(39), 13339–13342. https://doi.org/10.1523/JNEUROSCI.3814-12.2012

Wang, Y., Brzozowska-Prechtl, A., & Karten, H. J. (2010). Laminar and columnar auditory cortex in avian brain. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 12676–12681. https://doi.org/10.1073/pnas.1006645107

Woolley, S. C., & Doupe, A. J. (2008). Social context-induced song variation affects female behavior and gene expression. *PLoS Biology*, *6*(3), e62. https://doi.org/10.1371/journal.pbio.0060062

Woolley, S. M. N., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–1379. https://doi.org/10.1038/nn1536

Yost, W. A. (1986). Processing of complex signals and the role of inhibition. In P. R. D. Moore B.C.J. (Ed.), *Auditory Frequency Selectivity* (pp. 361–370). Boston, MA: Springer. https://doi.org/10.1007/978-1-4613-2247-4

Zann, R. (1996). Vocalisations. In *The zebra finch: a synthesis of field and laboratory studies* (pp. 196–247). Oxford University Press.