ABSTRACT

Title of Dissertation: Perceptual Binding and Temporal Coherence in the Auditory Cortex

Kelsey J. Dutta
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Shihab A. Shamma
Department of Electrical and Computer Engineering

Auditory streaming and perceptual binding are functions performed by the auditory brain rapidly and without conscious effort. They are fundamental for how we analyze and understand the sound environment including the perception of speech and ability to attend to one speaker while ignoring background noise. Recent work has suggested that temporal coherence of frequency components is a key cue which causes the brain to group channels and into a unified auditory stream. Coherent frequency inputs will lead to coherent neuronal firing, and we hypothesize that such neurons will demonstrate reciprocal enhancement of firing rate or suppression of responses to incoherent channels. This dissertation examines neuronal activity from the auditory cortex of ferrets in order to better understand the role of temporal coherence in formation of auditory streams. One experiment examines the role of temporal coherence in a selective attention task paradigm, and the other uses a stochastic figure-ground stimulus to examine neural correlates of a perceptual "pop-out" during passive listening. A third project develops a biophysically plausible model for a pitch-processing neuron in the early auditory system.

Perceptual Binding and Temporal Coherence in the Auditory Cortex

by

Kelsey J. Dutta

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Shihab A. Shamma, Chair/Advisor
Professor Behtash Babadi
Professor Jonathan Z. Simon
Professor Matthew J. Goupell
Professor Nikolas Francis

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever. First and foremost I'd like to thank my advisor, Professor Shihab Shamma for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past four years. It has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to thank and acknowledge the NSL lab research faculty and staff who have been instrumental along the way. Dr. Pingbo Yin provided valuable practical advice as well as stimulating conversations about scientific design and interpretations. Dr. Kai Lu provided understanding mentorship as well as patient technical training. Dr. Jonathan B. Fritz was an inspirational influence with his encyclopedic knowledge of the research canon and unbounded curiosity for all things neuroscience. Drs. Rupesh Chillale, Ali Mohammed, Diego Elgueda, and Dani Duque-Doncos provided a sounding board, technical advice, and helping hands along the way.

My colleagues at NSL have enriched my graduate life in many ways and deserve a special mention. Dr. Mohsen Rezaeizadeh was a constant cheerful presence and partner in commiseration. Dr. Neha Joshi has been an invaluable resource for work troubleshooting, a rare voice of sanity, and all-around an excellent friend. My former roommate Anna Lee O'Dell provided a needed reminder that real people outside of academia are impacted by our work.

My partner Garrett has put up with a lot of stuff over the years. Thanks, bub.

Finally, I would like to acknowledge financial support from the Center for Comparative Biology of Hearing (NIH T-32) and the Computation and Mathematics for Biological Networks training program (NSF institutional training grant).

Table of Contents

List of Figures

List of Abbreviations

A1	Primary auditory cortex
AC	Auditory cortex
AVCN	antero-ventral cochlear nucleus
CF	Center frequency
CL	Coherence Level
CN	Cochlear nucleus
BF	Best frequency
F0	fundamental frequency
LIN	lateral inhibitory network
LSO	lateral superior olive
MSO	medial superior olive
MGB	medial geniculate body
PSTH	peri-stimulus time histogram
PVCN	postero-ventral cochlear nucleus
IC	inferior colliculus (IC)
RF	receptive field
SFG	stochastic figure-ground
STRF	Spectrotemporal receptive field
TC	temporal coherence
TORC	temporally orthogonal ripples combination
VNLL	Ventral nucleus of the lateral lemniscus

Chapter 1: Introduction and Background

Like pixels in an image, sound enters the brain through sharply focused channels, which are interpreted to create our perceptions of the world. The auditory system creates internal representations of voices, music, and traffic noise from sound mixtures overlapping in time and frequency. The ability to hear out one voice in a crowded room is known as the "cocktail party problem," and is a long-standing mystery in auditory neuroscience [1]. However, an underlying question is how the brain selects and groups the composite frequencies of that voice into the unified percept of a single sound. In an acoustically cluttered environment, the auditory brain receives many overlapping frequency inputs, groups them into constituent objects, and then allows us to focus on one of interest, all rapidly and with little conscious effort.

What cues does the brain use to group components together, or perceptually bind them? One common attribute of natural sounds is harmonic structure, the spacing of frequency components at regular intervals. Experiment I addresses the perception of pitch with a biophysically plausible model of how harmonicity might be detected and encoded in the early auditory system. Another likely factor is temporal structure and coherence, or the synchronicity of the different frequency channels that make up one sound. Experiments II and III seek to understand the role of temporal coherence in auditory stream formation in both behaving and passively listening ferrets. This will dissertation will discuss experimental design, analysis and findings, and their

implications for hearing science and future investigation.

1.1 The Auditory Brain

1.1.1 Sensory inputs, the periphery to the midbrain

Sound enters the ear in the form of pressure waves which vibrate the eardrum, inducing movement of the basilar membrane in the cochlea. This causes hair cells along its length to depolarize and thus transforms a mechanical wave into an analog, electrical signal. The hair cells that depolarize in response to a given sound broadly correspond to its frequency components. The hair cell outputs pass through a biologically-implemented bank of bandpass filters followed by a lateral inhibitory network (LIN) in order to produce a high-resolution labeled line system of frequency tuning in the brainstem [2], see figure 2.1. In the auditory nerve, neurons with center frequency (CF) less than 1 kHz fire in phase with the incoming sound at that frequency. Most neurons cannot fire at a rate greater than 1 kHz, however normal-hearing adult humans can hear up to 18 kHz as the result of high-frequency channels responding to sounds in the unresolved frequency range [3, 4]. For example, a neuron with CF 10 kHz has a bandwidth of 1 kHz. While this neuron will respond to a 10 kHz pure tone, it will fire most predictably when multiple input frequencies fall in its bandwidth. For instance, if it receives 9.8 kHz and 10 kHz, it will beat at the difference frequency of 200 Hz. The power of this beating is enhanced when there are more harmonically spaced components, e.g. if 10.2 kHz is added. Notably, the different frequency inputs must be in phase to induce beating [5, 6]. In Experiment I (see Section 2), we focus on the outputs of the cochlea and LIN to propose a model of how the perceptual fusion of harmonic complexes might arise in the early auditory system.

After the cochlea, auditory signals are integrated in a series of structures in the brainstem. The neural encoding of sound undergoes filtering and transformations for various perceptual and behavioral functions (e.g. source localization), passing through the lateral lemniscus, inferior colliculus, thalamus, and arriving in the auditory cortex (AC). As signals ascend through the auditory system, neuronal response latencies increase and maximum firing rate decreases. Downstream from the midbrain, responses are no longer phase-locked to the stimulus however frequency specificity of responses, or tonotopic organization, is preserved through the auditory cortex.

Due to the difficulty of recording from the auditory midbrain, little is known about its function during awake listening or behavior. It is worth noting that in addition to the medial geniculate body (MGB) of the thalamus, the auditory midbrain includes the inferior colliculus (IC), a complex structure with extensive connections both to and from AC. Fields in IC exhibit spectral and temporal modulation tuning, which is inherited by AC [7]. Thus, theories of cortical function should keep in mind the feedback and feed-forward connections with the midbrain.

1.1.2 The auditory cortex

Auditory cortex (AC) is divided into three highly connected subfields (primary-tertiary), which are further parsed depending on anatomy, function (i.e. physiology), and/or species. The primary auditory cortex (A1), which is present in humans and higher order mammals, receives direct inputs from the thalamus, other auditory areas, and non-sensory areas associated with executive function, such as frontal cortex. While receptive fields (RFs) of auditory nerve cells can be summarized by 1-dimensional frequency tuning curves:, neurons at the midbrain [7] and cortex

are tuned to spectral and temporal modulations and are therefore described using spectrotemporal receptive fields (STRFs). Neurons in AC display frequency tuning often with one or more excitatory regions flanked by inhibitory sidebands. In A1, excitatory neurons exhibit temporal response diversity and a characteristic frequency tuning while inhibitory units are more temporally uniform but display a broader range of spectral properties [8].

AC is implicated in focusing auditory attention and neurons across fields exhibit rapid attention-dependent plasticity, that is they respond preferentially to sounds the listener is actively attending to, while responses to unattended distractor sounds are relatively suppressed. Numerous studies have shown that neuronal response properties are rapidly modulated by attention, enhancing the neural representation of pure tones [9, 10], multi-tone chords [11], and speech [12]. This modulation corresponds to the auditory stimulus and task difficulty, and STRF changes can persist for minutes to hours after the training session [9, 13].

Since the cortex lies directly beneath the skull, its activity can be measured via non-invasive recording methods such as electroencephalography and magnetoencephalography. It is also easily accessible to invasive methods such as electrocorticography and traditional penetrating electrodes.

1.1.3 Animal models for understanding human brain function

Many animal species perform stream segregation in the wild and some have been trained on streaming tasks (for review, see [14]). Cai and Dent [15] found that birds trained to selectively attend to a target stream were able to perform in the presence of a background stream, although increasing the salience of background sounds degraded performance. Animals are also able to

perform stream binding of ABAB sequences; Christison-Lagay and Cohen [16] showed that monkeys trained to indicate hearing one or two streams, were more likely to report two streams for alternating tones with larger Δ f, similar to human perceptual trends.

1.1.4 Justification of ferrets as the chosen model

Ferrets as an animal model have a long history in auditory neuroscience. Their hearing range is very similar to humans', they have folded brains with similar anatomical markers to us, and they are intelligent natural predators with the ability to learn complex tasks [17]. Ferrets have proven themselves capable of performing streaming with similar behavior to humans on alternating tone tasks [18], detecting a repeated noise stream mixed with a noise masker [19], and target detection with competing talkers [12].

1.2 Auditory Streaming

Human listeners quickly and easily tune in to sounds of interest and tune out distractors, even in noisy environments. This task of sound source segregation has been an open area of investigation since it was first described by Cherry in 1953 [1, 20]. The computational difficulty of segregating two or more sources is underscored by the many frequency components that make up even the simplest natural sounds. These components are often interleaved over the frequency spectrum and therefore cannot be separated by a simple low- or high-pass listening strategy. How the auditory system groups diffuse frequency components to create the perception of one or more sources is an open mystery.

Auditory objects are poorly defined but roughly analogous to sound sources, or the per-

ceptual model of sound sources. Griffiths and Warren [21] define auditory objects through their analysis and perception, situating them as internal representations not necessarily faithful to external realities. Since sound objects have a temporal dimension (unlike visual objects), a series of related sound events can be called an auditory stream. In his authoritative text on auditory perception, Bregman [20] defined a sound stream as a unified percept derived from a clustering of qualities such as timbre, loudness, and pitch. These qualities can only be experienced as properties of a stream in the same way that color, shape, and symmetry can only be experienced as properties of visual objects. Here, the terms auditory object and stream are used interchangeably.

Griffiths and Warren [21] outline three essential principles, or steps, of auditory object analysis— using sensory inputs from the external world, separating specific object-related information from other sources, and abstracting sensory inputs so an object can be generalized between separate events. This dissertation is mainly concerned with the middle step. Since sound information undergoes frequency analysis before entering the brain, the main challenge of sound perception is appropriately assigning constituent frequencies to form streams. This process of assignation is referred to as perceptual binding.

While it is enticing to frame the cochlea-to-cortex pathways as a multirate filter bank with near-perfect reconstruction, cochlear frequency analysis is blurred, biased, and subject to damage. The physiological limits of the cochlea result in imperfect isolation and frequency analysis of the sound waveform, thus the synthesis side must do much more than reverse the peripheral filter operations [22]. Fortunately, this incomplete information can be sufficient to generate an internal model of the source [23], and to estimate, and fill in occluded or unavailable information. Thus the relationship of the perceptual stream to the acoustic emanation of a source is nebulous and context-dependent.

Another useful but ill-defined concept in auditory scene analysis is that of an auditory texture. Sound textures are a class of sound that are usually considered backgrounds to speech or other salient streams. Saint-Arnaud and Popat [24] define texture as a category distinct from speech or music in that the latter's "potential information content" increases over time, while a sound texture has structured or random components that are stable over a long period and can be accurately estimated with a short sample window. Interestingly, a critical mass of streams can merge into a texture. For example, a dripping faucet is a stream but rain is a texture. Similarly, a few vehicles are streams but enough cars will be described as the texture of traffic. Saint-Arnaud and Popat [24] suggest that there is a maximum time between events, below which threshold individual but acoustically similar sources become indistinct and blur into a texture. By design or ecology, auditory textures are less salient than streams and form the background of naturalistic auditory scenes.

1.2.1 Pitch and streaming

Perceptual pitch is a potential low-hanging fruit for stream formation and segregation; harmonically spaced inputs readily undergo perceptual fusion by the auditory system, to the point where it is difficult to segregate two streams with octave-spaced fundamentals. The auditory system seems to have evolved to efficiently process harmonics without conscious effort. The musical consonance of tones with fundamentals related by small fraction, such as 5:4 and 3:2 corresponding to the major 3rd and perfect 5th of a major chord in Western music, is likely related to the two notes sharing half of their overtones. Similarly, smaller fractions such as 9:8 (2nd), 6:5 (minor 3rd) rarely overlap and create musical dissonance. Despite decades of intensive psychoacoustic

and physiological experimentation, the underlying neural substrates of pitch perception remain essentially unknown, and the necessary computations are at best ambiguous.[6].

Pitch processing is so sensitive that mistuning one harmonic causes the listener to perceive an additional stream; perceptual studies from the late 20th century demonstrated that mistuning one harmonic of a complex tone by as little as 2% cause that frequency to be heard as a separate tone [25, 26]. Popham et al. [27] extended this result to mistuned harmonics in speech, resulting in the percept of a concurrent whistle. Along the same lines, a harmonic target tone embedded in a noise is easier to detect than an inharmonic one [28] and concurrent, inharmonic streams are difficult to segregate [27]. In fact, noisy speech is represented "cleanly" (i.e. without noise) by the auditory cortex in naive, passively listening animals [29], suggesting that harmonic structure is salient across species and that pitch rescues streams from "fall[ing] apart into a sea of individual harmonics" [27].

As a caveat to these results, other studies have shown that harmonics are more useful for suppression than enhancement of an auditory stream; a pitched sound is easier to ignore [22]. Inharmonic speech alone is well intelligible, but significantly degrades performance when competing with a harmonic target vowel [30, 31]. de Cheveigné [22] summarizes that pitch facilitates stream segregation by assisting in harmonic cancellation of the background, rather than selective boosting of foreground. This phenomenon seems to extend even to the cancellation of many concurrent, pitched streams. Popham et al. [27] demonstrated that harmonic and inharmonic speech intelligibility were affected equally by speech-shaped noise but speech babble degraded inharmonic speech intelligibility more. They hypothesized that speech-babble allows spectral glimpsing of individual voices, introducing more ambiguity for spectral grouping. Thus, harmonic cancellation acts at multiple scales.

While it has been shown that AC encodes [32–34] and tracks [35] pitch, harmonic structure is not required for stream formation. For this reason, we intentionally avoid harmonic relations between components when designing stimuli to investigate temporal processing.

1.2.2 Temporal regularity

The auditory system is particularly sensitive to temporally regular sounds, as demonstrated by many behavioral results [36, 37] . Independent of conscious attention, mechanisms such as mismatch negativity alert the listener to violations in acoustic patterns, creating a potential confound for streaming experiments that rely on deviance detection. A common element of early streaming tasks is a temporally regular or periodic structure in the attended stream. Periodicity or pseudo-periodicity is a feature of many natural sounds, including speech and likely assists in target detection by targeting attentional resources, or temporally orienting attention [38]. Chait et al. [39] demonstrated that priming subjects to attend or ignore different time points can modulate cortical onset responses. However, temporal regularity does not result in improved performance in a selective attention gap detection task [40]. Similar to harmonic structure, periodicity may be more useful for ignoring [41] background than enhancing a foreground.

1.2.3 Temporal coherence and the cocktail party problem

Recent studies have suggested that temporal coherence (TC) is a key cue for stream formation. Many common sounds such as human and animal vocalizations contain temporally synchronized onsets. Elhilali et al. [42] proposed that an auditory stream is formed when the brain detects and groups spectral components that vary coherently together over time. To segregate, two

streams must activate different sets of peripheral filter channels. Thus, TC relies on synchrony of channels and spectral separability of sound sources.

This hypothesis was developed by Krishnan et al. [43] into a computational model for how streaming might occur using long time-scale coincidence detectors on the outputs of spectrotemporal modulation filters. These filters represent stimulus encoding in A1. The TC model reproduces human perceptual limits of streaming; Christiansen et al. [44] demonstrated that it can reproduce tone fission and TC boundaries of human perception at similar Δf and tone repetition times. Importantly, the TC model relies on onset synchrony rather than energy overlap. The model fused tones with onset asynchronies up to 20ms into a single stream for longer (75 ms) tones, and with asynchronies up to 10 ms for shorter (30 ms) tones. Contrary to earlier results by Bregman and Pinker [45], which suggest that fusion requires 88% or greater overlap, Christiansen et al. [44] suggest that streaming depends more on onset timing than temporal overlap. This is in line with previous results showing segregation of complex tones with asynchronous onsets but synchronous offsets, and therefore 100% overlap of the shorter tone [46].

1.3 The role of attention in streaming

Shamma et al. [47] related the TC hypothesis to an object-based attention theory of auditory scene analysis in which stream formation is pre-attentive, but stream selection is attention-dependent. Attention is then required to selectively enhance or suppress objects within the environment according to task demands. A key challenge of this theory is that many objects can be subdivided for focus into smaller objects, which can lead to potentially many streams-within-streams whose representations would be concurrently maintained by the auditory brain. For

example, it is unlikely that the brain maintains a representation of each instrument in an orchestra as a separate stream.

Selective attention is defined as attention to a single stream or source in the presence of masking sound(s). Subjects are often asked to report a target word or phrase to assess whether they successfully heard out the intended voice. Numerous studies have demonstrated that selective attention to a speaker results in relative enhancement of brain responses to the envelope of the attended stream and suppression of distracting noise [48] or competing streams [49, 50]. Other studies have shown individual components of competing streams are enhanced or suppressed according to attentional state [12, 51]. In passive animals listening to a speech mixture, both male and female speakers were represented in the neural response. In animals performing a female voice target detection task, however, AC responses showed strong relative enhancement of the female voice harmonics and suppression of the male's. This result is consistent with known dynamics of frequency-specific, attention-dependent plasticity in the ferret AC [9–11].

In contrast to selective attention, global attention listening tasks are designed to require attention to the sound scene as a whole. This is sometimes framed as auditory scene analysis [20], however the distinction between scene analysis and stream segregation is unclear; cocktail parties are scenes, too. In global attention tasks, subjects are usually asked to detect and report hearing a change, which can be applied to many different parameters of the sound [36, 40, 52–54]. The neural processes underlying global attention are not well-understood and likely vary depending on the change parameter. However, results from Lu et al. [53] indicate that in some tasks, neural activity in global attentive listening can mimic streaming activity. They saw suppression of A1 cell responses to alternating tones and enhancement to synchronous tones during engagement in a global listening task. The enhancement of responses to temporally coherent tones mimics known

perceptual fusion phenomena and further supports that idea that stream formation does not rely on explicit attention [47].

1.3.1 Pre-attentive Streaming

Attention can be driven by external or internal processes, where the former depends on task-relevance and the latter can involve ecological or learned salience, internal state, and other endogenous factors. While explicit selective attention modulates cortical responses [9], recent studies have suggested that differences in the salience of mixture stimuli could drive attention-like processes in the absence of explicit instruction [19, 55, 56]. Here, we equate attention-like modulation with any short-term, within-session response modulation not explained by sensory adaptation (e.g. stimulus-specific adaptation), and not to be confused with long-term changes in encoding [57]. This phenomenon has been observed in passively listening humans and animals, whose AC responses to speech in noise strongly resemble responses to clean speech; responses to noise are suppressed [29, 48]. In another study, human listeners alternatively listened to noisy speech or performed a visual task while similar auditory stimuli were presented. The subjects performed poorly on questions about the speech content during the visual task (confirming that they were not actively attending to the sound) however, there was no significant difference in the AC neural responses [48]. Suppression of background noise occured independently of attentional state. These finding suggests that attention-like plasticity in A1 can be driven by inherently stronger salience of speech over background white noise.

Response modulation during passive listening can also be driven by less-obvious difference in stream salience. Neurons in ferret AC segregate repeating from non-repeating streams of

broadband noise during passive listening [19]. Here, the repeating stream is more salient than the non-repeating one. To distinguish this effect from simple habituation, Dahmen et al. [58] showed that conditioning with tone pairs can change A1 response properties, but pure tone conditioning does not. Associations between tones was a sufficiently salient stimulus to induce short-term response modulation.

Recent studies have shown that incidental learning of sensory information can occur during task performance, suggesting that statistical regularities in the stimulus are detected and stored subconsciously. Gabay et al. [56] demonstrated incidental learning of auditory categories in an audiovisual task in which only the visual categories were explicitly directed, but audio information conferred a task advantage. Zhao et al. [55] similarly showed that statistical regularities bias attention in the visual system, even when recognizing the pattern did not confer a task advantage. These two findings suggest that undirected, sensory systems monitor and learn stimulus regularities.

In contrast, Southwell et al. [40] did not find evidence that statistical regularity (periodic repetition) endogenously attracts attention in a global listening task. Listeners were asked to detect a silent gap in one of two concurrently presented streams, one regular and one irregular. Task performance was similar for the target in both streams. This result may be explained by the simultaneous presentation of a task-irrelevant auditory stream necessitating active suppression or ignoring; the prior examples showed incidental learning of information which did not compete with the task-relevant dimension.

Neural responses to the concurrence of equally salient streams is seen in the body of literature on mixed-talker speech when the subject is passively listening and brain response represents the full sound mixture (or attention switches too rapidly for detection). Joshi et al. showed that in

trained but passively listening ferrets, male and female speakers were simultaneously represented in AC. The same animals showed strong response to the female voice and extinguished response to the male voice during task performance [12].

1.3.2 Neural mechanisms of streaming

The mechanisms of attentional modulation of neural activity are largely unknown. Zion Golumbic et al. [59] suggest two distinct candidate mechanisms: gain control and selectivity. The former requires that unattended stimuli are represented (in attenuated form) throughout the processing hierarchy while the latter allows unattended information to be discarded. Electrocorticography data suggest that both are present, and that selectivity and amplitude modulation (AM) exist on a continuum of plasticity in the auditory cortex. Interestingly, neural activity at highly selective sites increased over a few seconds time course, while AM sites did not display such a buildup [59]. This result is consistent with human and animal data showing gradual modulation of receptive fields for streaming.

Chapter 2: Spectro-temporal templates unify the pitch percepts of resolved and unresolved harmonics

2.1 Abstract

Pitch is a fundamental attribute in auditory perception involved in source identification and segregation, music, and speech understanding. Pitch percepts are intimately related to harmonic resolvability of sound. When harmonics are well-resolved, the induced pitch is usually salient and precise, and several models relying on autocorrelations or harmonic spectral templates can account for these percepts. However, when harmonics are not completely resolved, the pitch percept becomes less salient, poorly discriminated, with upper range limited to a few hundred Hertz, and spectral templates fail to convey percept since only temporal cues are available. Here, a biologically-motivated model is presented that combines spectral and temporal cues to account for both percepts. The model explains how temporal analysis to estimate the pitch of the unresolved harmonics is performed by bandpass filters implemented by resonances in dendritic trees of neurons in the early auditory pathway. It is demonstrated that organizing and exploiting such dendritic tuning can occur spontaneously in response to white noise. This paper then shows how temporal cues of unresolved harmonics may be integrated with spectrally resolved harmonics, creating spectro-temporal harmonic templates for all pitch percepts. Finally, the model extends

its account of monaural pitch percepts to pitches evoked by dichotic binaural stimuli.

2.2 Introduction

Pitch is the basic attribute of sound that conveys musical melody and contributes to speaker identity and speech prosody. It also plays a critical role in enabling listeners to organize cluttered soundscapes into their constituent sources [20, 60–63]. Yet despite decades of intensive psychoacoustic and physiological experimentation, the underlying neural substrates remain essentially unknown, and the necessary computations are at best ambiguous. Part of the difficulty stems from the ambiguity of the pitch percept itself, which is commonly associated with a variety of stimuli ranging from single and complex tones, to rapidly modulated noise [3, 64].

We focus here on the pitch evoked by harmonic complexes of different fundamental frequencies. The low order harmonics ($<$ 10th) in such a complex are normally aurally fully- or partially-resolved, i.e. their responses are mostly segregated into different frequency channels in the cochlear output as illustrated by outputs up to about 2 kHz in Fig. 2.1(a). Perceptually, the pitch induced by such harmonics is salient, well-discriminated, with a perceived value corresponding to the fundamental of the complex regardless of whether the fundamental and other nearby harmonics are present or absent. When the harmonics are aurally unresolved ($>>$12th), they co-occur within the cochlear bandpass filters and interact producing a "beating" or amplitude modulations that evoke a pitch sensation corresponding to the modulation rate (channels beyond about 2.4 kHz in Fig. 2.1a). However, this "unresolved" pitch is less salient, and is poorly discriminated [4, 65–68]. Many other contrasts exist between these two pitches that reflect the underlying harmonic separation. For example, randomizing the relative phases among the harmonics leaves

the resolved-pitch unaffected, whereas it reduces the saliency of the unresolved-pitch as it distorts the envelope of the beating waveforms. Another key observation concerns the frequency range of these pitch percepts. Unresolved pitches are typically limited only to a few hundred hertz [69], a limit whose origin is uncertain [70–72], whereas the resolved-pitch is perceived up to an order of magnitude higher frequencies (a few kHz). This latter limit coincides well with the assumed limits of phase-locking on the auditory-nerve [73, 74] (exceptions are discussed later [5, 6]).

The differences between the resolved and unresolved pitch percepts have often led to debates and speculations that they may arise from distinct mechanisms that exploit either temporal or spectral cues. Favoring a unified approach, "temporal models" (Fig. 2.1(b)) focus on the periodic structure of the phase-locked responses to all harmonics, and propose computations that directly estimate them, thus unifying the extraction of the resolved and unresolved pitch percepts. These models (e.g. the Auto-Correlogram [72, 75, 76] and others [70, 71, 77]) can explain most of the perceptual properties of resolved and unresolved pitch, including the weak saliency of unresolved pitch and its poor resolution, and the limited range of pitch up to a few hundred hertz [67].

Spectral models (Fig. 2.1c) by contrast exploit the spectral pattern of the resolved harmonics by matching it against presumed internal harmonic templates [78, 79]. Since unresolved harmonics lack this spectral signature, these template models do not attempt to account for unresolved pitch, nor for the reasons why the upper limit of resolved-pitch coincides with that of the phase-locking of its harmonics (see Refs. 19 and 20). Consequently, to account for the full range of pitch percepts, spectral models have been supplemented by additional distinct (presumably temporal) mechanisms for extracting unresolved pitch. The two percepts are then unified centrally to deliver consistent percepts [65, 80, 81].

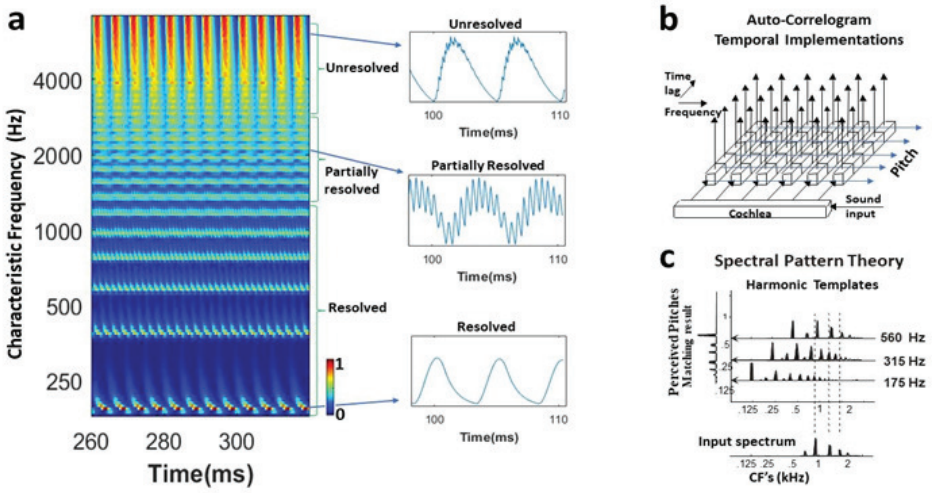


Figure 2.1: Harmonic resolution and models of pitch perception. (a) Auditory spectrogram of 40 harmonics of 200 Hz. Examples of fully resolved (200 Hz), partially-resolved (2200 Hz), and unresolved (7800 Hz) response waveforms are enlarged and shown to the right. As the harmonics become less resolved, they interact more causing the response envelopes to "beat" at the difference frequency. (b) The Auto-Correlogram model to measure the pitch of both resolved and unresolved harmonics. The autocorrelation function of each cochlear channel response is computed through a chain of temporal delays (lags). The functions are then summed across different channels to produce one combined correlation function whose peaks reflect the common pitch. (c) Spectral template matching to measure pitch of the resolved harmonics via harmonic templates. The inner-product of the incoming spectrum with each template produces a pattern of matches, whose peaks reflect the perceived pitches. Color bars next to each spectrogram indicate the scale of the response in arbitrary units.

In a previous study [79], we presented a view that intimately linked temporal and spectral cues in the formation of the harmonic templates. The model demonstrated how temporal phase-locking facilitates the spontaneous formation of the harmonic templates without need for training by explicit harmonic exemplars, and how the templates are subsequently used to estimate the resolved-pitch percept. It, however, did not integrate into these templates the computation of the periodicity from the unresolved harmonics modulations, and therefore left open the fundamental question of how this periodicity can become spontaneously linked to the spectral templates, i.e. how can a 100 Hz modulation rate be heard at the same pitch as that of the harmonic template pattern of a 100 Hz fundamental without explicit training to set up this correspondence?

This report attempts to fill in this gap by proposing a biologically-inspired model that builds upon and augments the harmonic templates model [79]. Specifically, our focus is on the unresolved pitch and its unification into the framework of the resolved harmonic templates [79]. We shall do this by explaining (1) how unresolved harmonic modulations can be readily analyzed, and their perceptual values estimated by dendritic processing; (2) how these computations become incorporated into the spectral harmonic templates through the same learning process, which spontaneously gives rise to the templates without need for harmonic exemplars. We shall then (3) discuss why the unresolved pitch range is limited, and (4) implement and simulate the model's outputs emphasizing stimuli that combine resolved and unresolved harmonics. Finally, we (5) offer a discussion of the physiological evidence of such a model, and ideas for how it might be tested, ending with (6) a brief account of how distinct binaural pitches are consistent with this model.

2.3 Methods

We summarize here the algorithms used to simulate the function of the proposed extended templates—referred to as the spectro-temporal templates—and how we utilize them to compute the pitch estimates from both resolved and unresolved harmonic inputs.

2.3.1 Cochlear frequency analysis and the auditory spectrogram

In many figures in this paper, we display an auditory "spectrogram" representation derived from a computational model of cochlear filtering followed by a LIN stage that exploits phase-locking in auditory-nerve responses to sharpen the spectral analysis and endow it with robustness to sound-level changes [2]. All aspects of this model have already been described in detail in several publications [82–85]. Briefly, cochlear analysis transforms sound through a series of stages in the early auditory system from a one-dimensional pressure time waveform to a two-dimensional pattern of neural activity distributed along the tonotopic (logarithmic) frequency axis, referred to as the auditory spectrogram. It represents an enhanced and noise-robust estimate of the Fourier-based spectrogram [84], but it differs in many respects due to its biophysical details as previously explained [82, 84, 85].

2.3.2 Mathematical formulation

The early auditory stages are illustrated in Fig. 2.2a [2]. The first operation is an affine wavelet transform of the acoustic signal s(t), which approximates the spectral analysis performed by the cochlear filter bank. This analysis stage is implemented by a bank of 128 overlapping constant-Q ($Q_{10dB} \sim 10$) bandpass filters with center frequencies uniformly distributed along

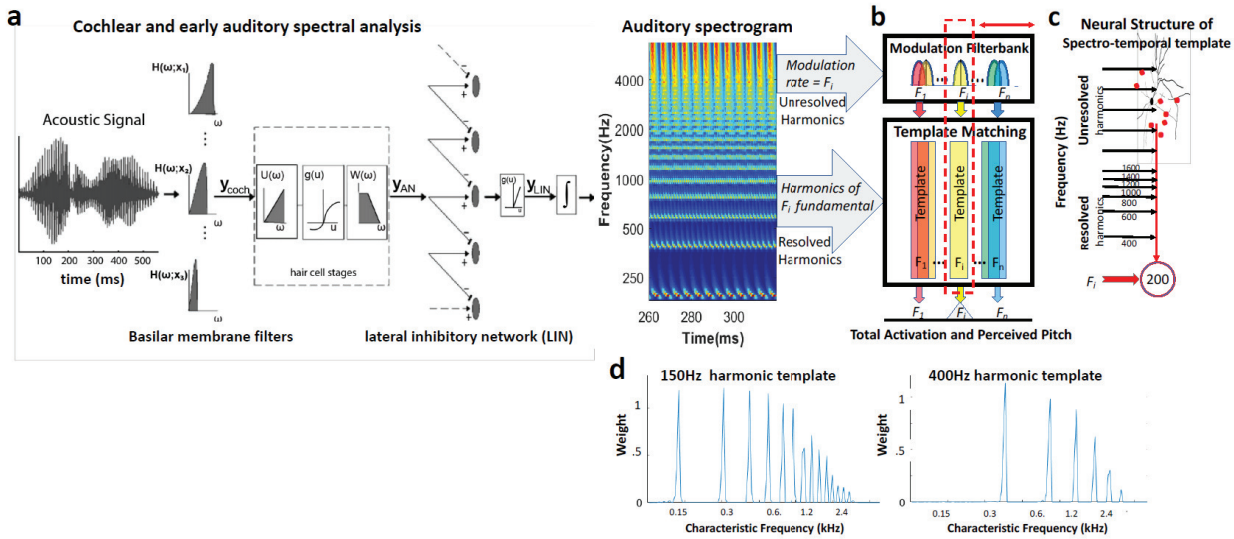


Figure 2.2: Schematics of the Computational stages in the spectro-temporal harmonics templates model. (a) Schematic of early auditory stages. The acoustic signal is analyzed by a bank of constant-Q cochlear-like filters. The output of each filter y_{coch} is processed by a hair cell model y_{LAN} followed by a LIN, and is finally rectified y_{LIN} and integrated to produce the auditory spectrogram y_{final}.(b) Schematic of the computations for the unresolved (top panel) and resolved (bottom panel). For the unresolved, periodicity estimation of the amplitude modulated cochlear inputs is done via a bank of bandpass filters. For the resolved components, a harmonic-template matching algorithm is applied [79]. (c) A proposed biological implementation of the two pitch computations. (d) Two resolved harmonic templates: (left) 150 Hz-fundamental and (right) 400 Hz-fundamental templates. Synaptic weights are applied to each cochlear channel, with peaks at locations harmonic to F0.

a logarithmic frequency axis (x), over 5.3 oct (24 filters/octave). Each cochlear filter is implemented by a minimum-phase impulse response function h(t;x) with a magnitude frequency response

$$H(x) = \begin{cases} (x_h - x)^{\alpha} e^{-\beta(x_h - x)} & \text{if } 0 \leq x \leq x_h \\ 0 & \text{if } x > x_h \end{cases} \tag{2.1}$$

where x_h is the cutoff frequency, α = 0.3, and β = 8 [86]. The cochlear filter outputs $y_{coch}(t,x)$ are transformed into auditory-nerve patterns $y_{AN}(t,x)$ by a hair cell stage consisting of a high-pass filter, a nonlinear compression g(·), and a membrane leakage low-pass filter w(t) accounting for decrease of phase-locking on the auditory nerve beyond about 2 kHz. The final stage simulates the action of a LIN postulated to exist in the cochlear nucleus [87], which effectively enhances the frequency selectivity of the cochlear filter bank [83, 88]. The LIN is approximated by a first-order derivative with respect to the tonotopic axis and followed by a half-wave rectifier to produce $y_{LIN}(t,x)$. The final output of this stage is computed by integrating $y_{LIN}(t,x)$ over a short window, $\mu(t;\tau) = e^{\frac{t}{\tau}} u(t)$, with time constant = 0.5 ms mimicking the further loss of phase-locking observed in the midbrain. The mathematical formulation for this model is then summarized as follows:

$$y_{coch}(t,x) = s(t) \otimes_t h(t;x), \tag{2.2}$$

$$y_{AN} = g(\delta_t y_{coch}(t,x)) \otimes_t w(t), \tag{2.3}$$

$$y_{LIN} = max(\delta_x y_{AN}(t,x), 0), \tag{2.4}$$

$$y_{final}(t,x) = y_{LIN}(t,x) \otimes_t \mu(t;\tau), \tag{2.5}$$

where $\otimes_t$ denotes convolution operation in the time domain, and δ_t, δ_x are derivatives with respect to time and cochlear axis. The model described above captures several of the important properties of auditory processing that are critical for our objectives, but it is highly simplified and lacks many of the known details of cochlear processing that are likely to be important in certain circumstances and should be added when needed [89]. All these issues are discussed in detail in [2]. For Matlab implementations, see [90].

2.3.3 Models of synaptic plasticity at the spectro-temporal templates

The spectro-temporal templates embody two sets of computations as depicted in Fig. 2.2b: (i) bandpass filtering of unresolved harmonics (top panel), and (ii) harmonic pattern matching for the resolved components (bottom panel). The full spectro-temporal template is envisaged to be a neuron as depicted in Fig. 2.2c; it is referred to later as the "pitch-neuron" for reasons that will be elaborated upon further later in the text. Both resolved and unresolved harmonic inputs from the auditory spectrogram innervate the pitch-neuron and form the appropriate connectivity patterns spontaneously with no supervision from ideal exemplars. The resolved harmonic input synapses

are formed as described in detail in a previous publication [79]. Thus, we shall focus here on describing the same basic adaptive process that connects the unresolved harmonic inputs into the dendritic region of the pitch-neurons, and results effectively in the formation of a bandpass filter centered at the characteristic frequency (CF) of each pitch-neuron.

We begin by assuming that the cochlea is driven by broadband noise. Each pitch-neuron is assumed to be driven by a CF cochlear input that initially induces an intracellular potential that is phase-locked to it. Other contributions to the intracellular potential might be from any other resolved harmonic inputs that have already formed (as schematically depicted in Fig. 2.2c). These inputs induce phase-locked spiking in the pitch-neuron. The spikes in turn produce intracellular potentials that are phase-locked to them (i.e. at CF), and that backpropagate up the neuron's dendritic tree. These potentials become attenuated on their way up the dendrites by pathways that are tuned (or resonant) to various frequencies that do not match those of the phase-locked backpropagating signals (at the CF of the neuron). This array of dendritic resonances arises from the anatomical and electrical structure of the dendrites as discussed in the text. Here, it is computationally simulated by a bank of bandpass filters tuned at all frequencies up to about 800 Hz, a range that exceeds the resonance frequencies that are typically found along dendritic locations [91].

To simulate synapse formation, we compute the coincidence between the incoming noisy auditory spectrogram and the "backpropagating" dendritic intracellular potentials of the pitch-neurons. For example, Fig. 2.4c depicts for each pitch-neuron (whose CFs are arrayed along the x axis) the strength of these coincidences, which in turn result in formation of synapses (indexed by their dendritic resonances along the y-axis). Thus, the pitch-neuron tuned at CF = 200 Hz (dashed black line in Fig. 2.4c) will form strong dendritic synapses only at locations tuned to 200

Hz, and a few harmonic multiples of that. The same occurs at all other pitch-neurons, a learning process that is elaborated upon further in the text.

2.3.4 Monaural pitch estimation

Monaural pitch estimates are made as depicted in Fig. 2.2b from both the resolved harmonics (using harmonic templates) and the unresolved harmonics (via the array of dendritic bandpass filters). For the resolved portion, two examples of resolved harmonic templates $W^k(f)$ are shown in Fig. 2.2d: W^{150} = 150 Hz and W^{400} = 400 Hz. The templates act a sieve to weight the array of cochlear spectral channels (S_f) at the corresponding harmonic locations. The net output from each template is the waveform

$$O^k(t) = \sum_f S_f(t) W^k(f). \tag{2.6}$$

Note that this output will be phase-locked to the composite waveform from all inputs. To estimate the overall power in the response, the waveform is half-wave rectified ($g(\cdot)$) and its root-mean-square (rms) power estimated as

$$P^k_{resolved} = \sqrt{\sum_t g(O^k(t))^2 / T.} \tag{2.7}$$

2.3.5 Binaural pitch estimation

We propose several possible schemes for binaural interaction of monaural pitch estimates. All approximately produce the results depicted in Figs. 2.7(a) and (c).

Simple summation

$$P^k_{unresolved} = \sqrt{\sum_t (P^k_{left})^2/T} + \sqrt{\sum_t (P^k_{right})^2/T}, \tag{2.8}$$

or simple lateral superior olive (LSO) subtraction and MSO summation (Fig. 2.7(b), top)

$$P^k_{LSO} = \sqrt{\sum_t (P^k_{left} - P^k_{right})^2/T}, P^k_{MSO} = \sqrt{\sum_t (P^k_{left} + P^k_{right})^2/T}. \tag{2.9}$$

Complex cross-inhibition (Fig. 2.7(b), bottom])

$$P_1 = P_{est,left} - P_{est,right}, \tag{2.10}$$

$$P_2 = P_{est,right} - P_{est,left}, \tag{2.11}$$

$$P_{xx} = \sqrt{P_1^2} + \sqrt{P_2^2} \tag{2.12}$$

2.4 Modulation Rates Due To Unresolved Harmonics

Speech, music, animal vocalizations, and many environmental and percussive sounds have a broad range of harmonics that extend, in human cochlear analysis, over both the resolved and unresolved ranges. As illustrated in Fig. 2.1a, lower resolved harmonics (<10th) induce localized responses that are phase-locked to their frequencies if less than 3–4 KHz. However, as the resolution deteriorates for the higher harmonics (>10th), they mutually interact within a cochlear filter, producing envelope modulations at the difference (fundamental) frequency of the complex (200 Hz in Fig. 2.1a). The depth of this modulation increases as the harmonics become more densely packed. Furthermore, the exact shape of the envelope depends on the relative phases of

the interacting harmonics, being strongly peaked if the harmonics are in-phase, and more variable if the phases are relatively random. The carrier of this AM waveform remains phase-locked near the center frequency of the cochlear filter if it is <3–4 kHz. Beyond this, the carrier is smeared out leaving only the envelope modulations (upper unresolved region in Fig. 2.1a).

As is well-known, subjects listening to the full harmonic complex report a salient pitch that is finely discriminable, and equal to the fundamental of the harmonic series, regardless of its presence or absence [64]. When listening only to the upper unresolved harmonics, subjects still report the same pitch value, but the percept becomes less salient and poorly discriminable. The pitch is weaker still if the relative harmonic-phases are randomized. It is important to appreciate that these envelope modulations can occur at any location along the tonotopic axis where the unresolved harmonics are found. For example, for the 200 Hz harmonic complex in Fig. 2.1a, the 200 Hz envelope modulations become substantial beyond about the 10th harmonic (2 kHz). Consequently, unlike the well-resolved lower harmonics which are found at specific frequency channels (200, 400,... Hz) and hence can be associated with specific fixed harmonic templates, the unresolved pitch is perceived from the envelop modulations wherever they occur. This also implies that the same frequency channels (e.g. near 2 kHz) may exhibit envelope modulations at arbitrary rates depending on the unresolved harmonics within the cochlear filter. For example, if the fundamental frequency is <200 Hz, then the channels near 2 kHz would exhibit strong envelope modulations at whatever the fundamental frequency is.

From a computational point of view, a simple scheme to measure these modulation rates, and hence account for the perceived unresolved pitch, is to use a bank of bandpass filters for each cochlear channel (Figs. 2.2b and 2.3), whose outputs are then summed to generate the overall unresolved pitch value. This scheme is essentially equivalent to the Auto-correlogram model

[75, 92] alluded to earlier (Fig. 2.1b), where the delay-lines of the autocorrelation effectively implement the bandpass filters. The same idea is also effectively implemented by the delays proposed in Ref. [79] (Fig. 2.1(b)). Physiological experiments have demonstrated rate-selective modulation responses analogous to bandpass filtering [93–96], but it remains unclear how these filters exhibit or account for the limited range or poor resolution of the unresolved pitch percepts, and how they can be unified with the resolved pitch percept.

2.4.1 Dendritic resonances as bandpass filters

Neurons propagate electrical currents from their dendritic trees to the soma, where they induce axonal spiking that represents the integrated dendritic inputs. Neuronal topology, structure, and active membrane channels have long been known to result in transfer functions between any point on the dendrites and the soma that resemble a bandpass filter, in that it is tuned to a specific frequency that reflects many biophysical parameters. A detailed study of such "bandpass resonances" in a typical dendritic tree (depicted in Fig. 2.3b) [91] revealed several important findings that have direct implications for the bandpass filter bank idea needed for unresolved pitch computations. First, in any given dendritic tree (Fig. 2.3b), synaptic inputs at different locations are filtered differently depending on the unique dendritic topology and the details of the path from each synapse to the soma. This means that there exists a range of bandpass filters tuned to different frequencies. Based on a detailed study of such trees [91], it has been found that their basic geometry and structural constraints limit the range of resulting bandpass filter tuning frequencies to a maximum of a few hundred hertz, and their bandwidths to be moderate (Q = 1). Therefore, if these dendritic bandpass filters are involved in the measurement of unresolved pitch, their limited

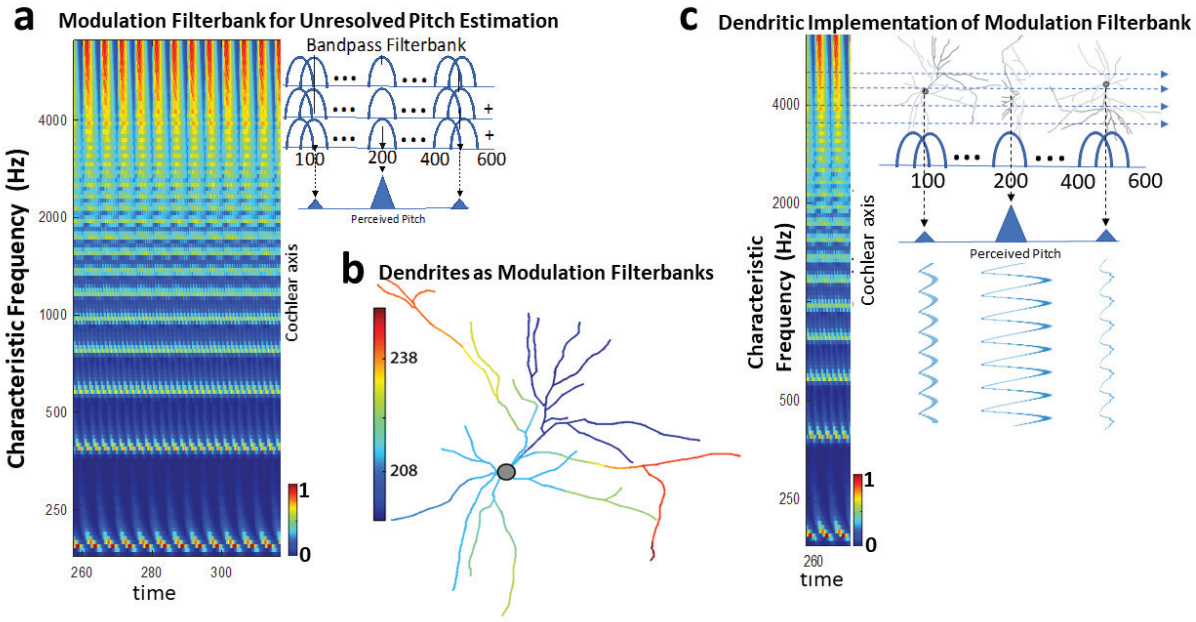


Figure 2.3: Analysis of temporal modulations on the auditory-nerve. (a) Each auditory-nerve fiber is schematically filtered by a bandpass filterbank tuned to a range of frequencies. The output from all filters is then summed (downward arrows) to estimate the final perceived unresolved-pitch. (b) Schematic of a dendritic tree as a bandpass filter bank (Figure 3.4 in [91]). The colors represent the tuning of the path between the synapses on these branches and the soma of the cell [91]. (c) Implementing the bandpass filterbanks with dendritic trees. Each neuron receives inputs from all auditory-nerve fibers whose synapses at a given dendritic tree are at locations that have paths to the soma tuned only near one modulation-rate. Thus, there is a neuron whose auditory-nerve inputs are filtered near 100 Hz; others are tuned at 200, 400, ... In the schematic, the most responsive dendritic tree is the one tuned to 200 Hz because the auditory-nerve modulations are at 200 Hz. Color bars next to each spectrogram indicate the scale of the response in arbitrary units.

range and poor resolution can readily explain the properties of this percept.

Figure 2.3(c) illustrates schematically how dendritic bandpass filters can serve to estimate the modulation rates induced by unresolved harmonics on the auditory nerve. We first assume that the dendritic trees of a certain type of neurons (we shall refer to as "pitch-neurons") are initially innervated by auditory inputs from a wide range of cochlear tonotopic locations. To make each pitch-neuron selectively responsive to a single modulation rate only, we shall describe a simple learning (or adaptive) process though which all inputs (synapses) are discarded except those arriving to locations tuned to the same rate. Thus, such a pitch-neuron becomes effectively tuned and responsive only to one modulation-rate. Different pitch-neurons would have dendritic inputs tuned to different modulation rates, effectively creating the bank of bandpass filters needed for unresolved pitch estimation. In Sec. 2.4.2, we discuss how this organization naturally emerges without specific training on harmonic exemplars. We shall subsequently address (in Sec. 2.4) the key question of how these rate-selective dendritic inputs are integrated into the harmonic templates needed to measure the resolved pitch, thus giving rise to the array of pitch-neurons that measure the complete pitch percept.

2.4.2 Shaping dendritic rate selectivity

How does the dendritic tree of a pitch-neuron become selective to one input modulation rate? Consider a neuron tuned to a particular CF and possessing an elaborate dendritic tree with auditory inputs from a wide range of cochlear locations (Fig. 2.4a). The CF is defined by the main input near the soma as indicated in the figure. We assume that the CFs of all such neurons to be $<$4 kHz (upper limit of phase-locking in most mammals), and hence the CF somatic inputs are

phase-locked. The precise upper limit of phase-locking is unknown, with estimates ranging from 1.5 kHz to 10 kHz, however the upper limit of pitch perception is around 4-5 kHz fundamental frequency [6]. Figure 3.4 (a) schematically depicts an array of such neurons with various CFs (200, 350, 700 Hz). We begin with an acoustic noise that excites the cochlea broadly at most of its tonotopic locations. Initially, the somatic CF input at each neuron induces a phase-locked intracellular potential that excites phase-locked spiking of the cell at the same CF rate (e.g. at 200 Hz), which in turn generates a correspondingly phase-locked signal that backpropagates up to the dendritic tree [97]. However, this phase-locked CF signal becomes attenuated in the dendritic tree at all branches except those tuned to the same CF. For example, the neuron with CF = 200 Hz would phase-lock its spikes to 200 Hz, and the resulting phase-locked intracellular potentials propagate up the dendritic tree, most strongly to dendrites whose pathway to the soma is tuned at 200 Hz [98–100]. So, if the input to these synapses from the cochlear channels is also phase-locked to 200 Hz, then the synapses may become enhanced because of their potentially correlated pre- or post-synaptic signals. However, if either is weak or unmatched in frequency to the other, then the synapse weakens [98, 99, 101].

We now consider the spontaneous formation of the rate-selectivity in each recipient neuron when its input is the cochlear responses to *white noise*. At all cochlear CF locations, auditory-nerve responses to the noise are modulated randomly with a bandwidth that reflects the range of frequencies interacting within the cochlear filters (Fig. 2.4a). We assume initially that the cochlear outputs projecting to the dendritic trees innervate all dendritic locations (indicated by the multicolored synapses in Fig. 2.4a). However, among all these inputs, only those at dendritic locations tuned near the CF of each recipient neuron would have a post-synaptic signal propagating from the soma (red synapses for CF = 200 Hz neuron, green synapses for CF = 350 Hz neuron,

and blue synapses for CF = 700 Hz neuron). Therefore, only these synapses become strengthened by the "Hebbian" correlation between pre- and post-synaptic potentials. At all other locations, the dendritic post-synaptic signal is too weak and hence the synapses are weakened and eventually pared down. In this manner, we postulate that the array of recipient neurons with different CFs become connected to cochlear inputs that are effectively bandpass filtered with dendritic filters tuned to the same frequency as the CFs. This final arrangement is illustrated for the neuron of CF = 200 in the left panel of Fig. 2.4b, where only the red synapses (tuned to 200 Hz) remain, while all others are lost. Such a rate-selective neuron is what we referred to earlier as a "pitch-neuron," which is tuned only to a specific modulation rate.

We simulated this learning process using a model of cochlear processing that included all major stages such as bandpass filtering, transduction, and phase-locking up to about 3–4 kHz [2,79] (see Appendix A for more details). Note that this early auditory processing model includes a lateral-inhibitory stage (LIN) [82,83], which extracts from the phase-locked responses of the auditory nerve a sharp and level-robust spectral representation [84,85], equivalent to cochlear tuning of Q $\sim$ 10, which has been justified for example in detail in Fig. 2 of Ref. [102]. Examples of such auditory spectrograms to noise and harmonic stimuli have also already been illustrated and explained in detail in Refs. [95] and [99]. These stages are schematically illustrated in Fig. 2.2(a).

Similarly, the pitch-neuron and its dendritic tree are modeled by the primary functional parts needed here to explain the learning and filtering process. For example, we represent the dendritic tree resonances by a bank of bandpass filters with moderate resolution (Q $\sim$ 1) as explained earlier based on the dendritic modeling [91]. Initially prior to the learning, the somatic signal is the cochlear noisy response at the neuron's CF. This signal causes spiking in the neu-

ron, which in turn induces intracellular potentials that are backpropagated through the dendritic bandpass filters whose outputs represent the post-synaptic potentials available in this neuron. Of course, the strongest surviving post-synaptic signal is that at the dendritic locations tuned near the CF (e.g. 200 Hz in Fig. 2.4b). To simulate the spontaneous formation and strengthening of the synapses, we computed for each pitch-neuron the integrated (over 2 s window) power in the coincidence between all dendritic post-synaptic potentials in this pitch-neuron and the pre-synaptic auditory spectrogram responses into the dendritic tree. When both the pre- and post-synaptic potentials contain equal frequencies (e.g. 200 Hz), then the average coincidence increases despite their random relative phases (exactly the same as the coincidences previously described in Ref. [79], which resulted in the selective inputs of the resolved harmonics). Figure 2.4(c) illustrates for each pitch-neuron (arrayed along the x axis) the amount of coincidence (or synaptic strengthening) at all of its dendritic locations, arrayed along the y-axis and indexed by the resonance frequency of each location. For example, in the neuron with CF = 200 Hz (marked with the dashed black line), the strongest formed synapses are those at dendritic locations tuned to 200 Hz, and weaker ones at the multiples of 200 Hz, all of which are highlighted by the white circles. The strong red trajectory highlights the strongest synapses formed, namely at the dendritic locations (y-axis) that happen to be resonant at the CF of the corresponding pitch-neuron (x axis).

To summarize, the dendritic resonances (Fig. 2.3(b)) perform the direct measurement of the unresolved-pitch modulations, much like the bandpass filters hypothesized in the schematics of Figs. 3(a) and 3(c). Thus, after learning and synapse strengthening, unresolved components of a harmonic complex at high frequencies generate modulated waveforms at the fundamental (e.g. 200 Hz), which best excite the pitch-neuron tuned at the same frequency (CF = 200 Hz), i.e. as in Fig. 2.4b (left panel). However, since dendritic resonances like those described above (and in

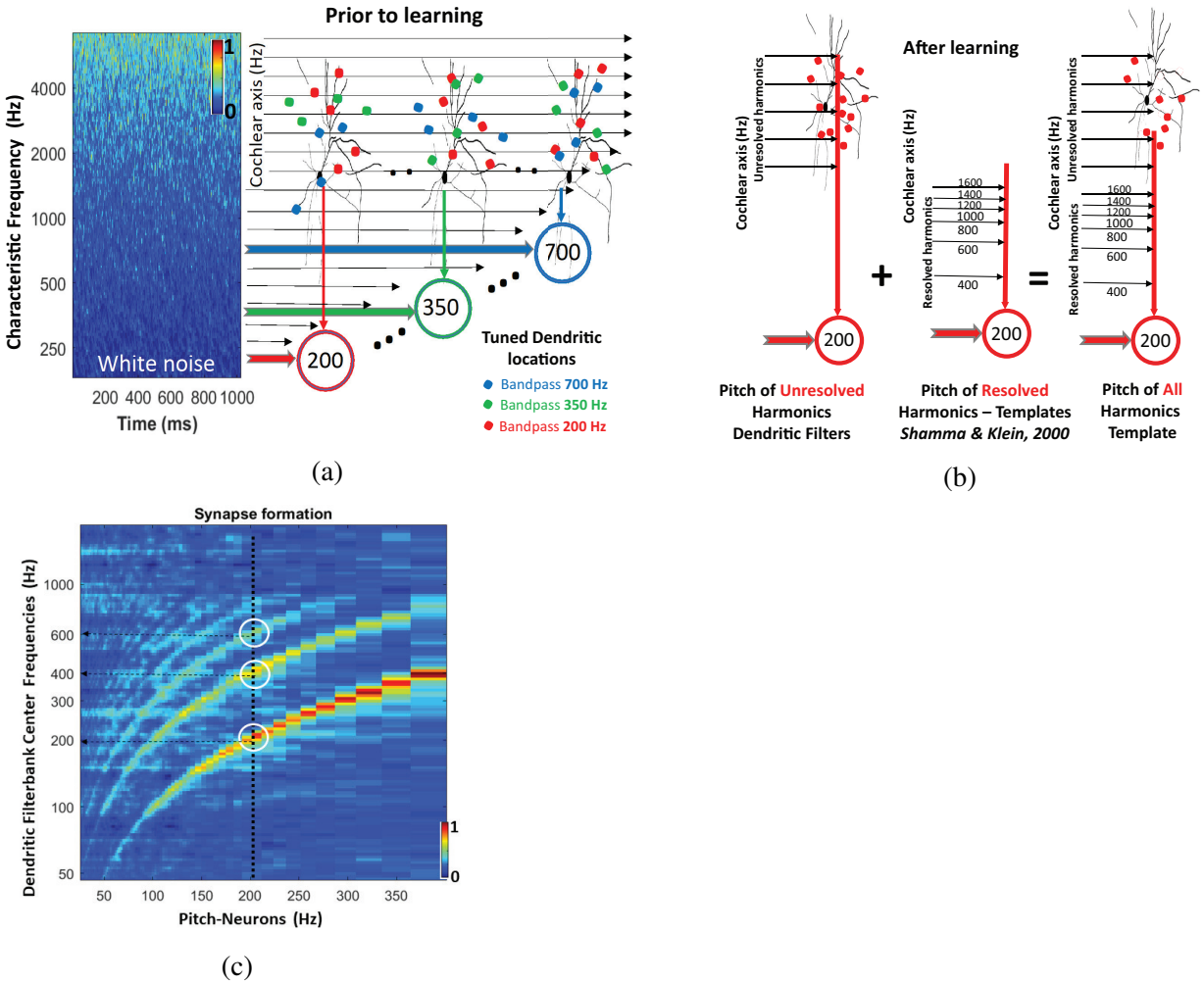


Figure 2.4: Learning the dendritic bandpass filters. (a) Prior to learning: Schematic of an array of neurons tuned at various CFs. The CF input to the neurons is indicated by the thick colored arrows (shown are CF's=200, 350, 700 Hz, as Red, Green, and Blue respectively). Initially each neuron has a dendritic tree innervated by a broad array of cochlear inputs. The synapse locations are symbolized by colored dots that reflect their tuning. Also shown are the cochlear responses evoked by a white noise input. (b) Spectro-temporal templates: (Left panel) Temporal modulations due to unresolved harmonics induce the formation of synapses at dendritic cites that are tuned to the CF of the cell. For example, the pitch-neuron here is tuned at CF=200 Hz. Only synapses at dendritic locations tuned to the same frequency (=200 Hz) persist, while others are eliminated (compare this to synapses prior to learning). In the end, the sole cochlear amplitude-modulated inputs that excite this neuron are those that are modulated at near its CF rates. (Middle panel) Resolved harmonic inputs spontaneously form inputs during the same learning process from inputs at harmonic multiples of the CF, forming the (classic) harmonic template of 200 Hz as described in [79]. (Right panel) Combining all resolved and unresolved harmonic inputs yields the complete pitch-neuron, referred to as the spectro-temporal harmonic template. (c) Simulating the learning: The average coincidence between the intracellular potentials and the noise-generated cochlear responses. It is selectively enhanced only at the dendritic inputs that are tuned at CF (x-axis). They connect to auditory inputs from all cochlear locations with filter bandwidths ¿ CF of the neuron (see text). Color bars next to each spectrogram indicate the scale of the response in arbitrary units.

detail in Ref. [77]) are largely limited to a few hundred Hz [77], then only pitch-neurons tuned to such low CFs can utilize the backpropagated somatic potentials in the dendritic trees to form the appropriately tuned synapses. Thus, although temporal modulations created by unresolved harmonics can exceed a few hundred hertz, these high modulation rates remain unperceived because of the lack of corresponding pitch-neurons that can measure them.

Next, we consider the frequency range in the transition between resolved and unresolved harmonics for each pitch-neuron. Pitch-neurons form resonant synapses with cochlear inputs from a wide range of CFs (Fig. 2.4b left-panel) as long as these input channels can provide the temporally modulated inputs (due to the unresolved harmonics). However, the lower limit of the CFs of these cochlear inputs is dictated by their bandwidths, which become progressively narrower at lower tonotopic locations, effectively providing responses to the resolved harmonics. For example, in Fig. 2.1(a), assuming roughly an effective cochlear tuning of 10% (i.e. Q = 10), cochlear outputs below about CF = 2 kHz (10th harmonic of 200 Hz) produce only weak temporal modulations at 200 Hz since they are partially resolved. Thus, cochlear channels below this CF will not form the resonant synapses on the dendritic tree of the 200 Hz pitch-neuron. Instead, this pitch-neuron would form the harmonically resolved synapses (the classic harmonic template) as described in detail in Ref. [79]. Therefore, as we discuss in more detail next, we recognize two regions of connectivity from the cochlea to each pitch-neuron as illustrated in Fig. 2.4b: the resolved harmonics region [103] (Fig. 2.4b, middle-panel) and the unresolved harmonic inputs (left-panel). The total combination of these inputs is what we refer to as the spectro-temporal harmonic template (Fig. 2.4b, right-panel).

2.5 Spectro-temporal templates: Physiological and Psychoacoustic Evidence

In a previous study, we examined how neurons similar to the pitch-neurons spontaneously form synaptic inputs that are harmonically related to the CF of the neuron [79]. To review briefly, the post-synaptic somatic potential due to a half-wave rectified and sharpened CF input can correlate with pre-synaptic inputs arriving at from cochlear channels tuned at multiples of the CF. These template connections arise regardless of the input acoustic stimuli, and even if the cochlea is driven by white noise. This is because the inputs arriving at CF, and from cochlear filters tuned to its low-order multiples (e.g. CF = 200 Hz, and multiples up to approximately the 10th harmonic) are (at least partially) resolved and phase-locked near their carrier frequencies. Consequently, significant coincidence (integrated over a long-duration window) will occur between the somatic CF potential and each of the signals arriving from cochlear filters at multiples of the CF (e.g. 2 CF, 3 CF). As we discussed above, when this signal propagates further up the dendritic tree, higher frequencies are filtered out and the backpropagated signal will only reach regions that are tuned to the CF or its multiples, if they are within a few hundred hertz. Synapses will then form at these locations if the cochlear inputs contain modulations at the CF frequencies, which they do for the noise stimulus discussed earlier (Fig. 2.4c). After learning, this pitch-neuron becomes connected and driven by the two kinds of cochlear channels (Fig. 2.4b), forming the "spectro-temporal harmonic templates" as we discussed above. Note that in all subsequent figures, each pitch-neuron is now labeled by a pitch value that is both the CF of the neuron (in Hz), or also the fundamental of the harmonic complex that evokes the best response among all pitch-neurons.

In order to illustrate how these pitch-neurons extract and convey the pitch percept, we il-

lustrate in Fig. 2.5(a) the average power in the responses of an array of such pitch-neurons tuned to a wide range of CFs, and forming an axis which would display a maximum indicating the pitch value by its location, and pitch saliency by its power. The resolved harmonics template used in this and all similar subsequent computations of the resolved harmonics pitch is shown in Fig. 2.2(d). The stimulus here is composed of a harmonic complex of 200 Hz. The output at each pitch-neuron is computed as detailed in Methods (Sec. 2.3); it consists of the superposition of the inputs coming into the neuron from the resolved (blue traces) and unresolved (green traces) inputs; their superposition is the total activation of the pitch-neurons (red traces). As expected, these activation patterns typically exhibit multiple other peaks that reflect the partial correspondence between different harmonic series. For example, 100 and 200 Hz harmonic series share many harmonics in common. These "simultaneous" peaks in Fig. 2.5 can be interpreted as concurrently heard pitches whose percept is largely fused with the main pitch (200 Hz).

Figure 3.5(b) illustrates the decrease in saliency of the unresolved-pitch when the harmonics change from in-phase (Fig. 2.5a) to random relative phases (Fig. 2.5b) causing the peak of the green trace to decrease in height (decrease in saliency). The resolved pitch percept is robust to phase-changes, and here it remains unaltered between the two conditions (blue trace).

Finally, since pitch-neurons are driven by phase-locked cochlear responses both for the resolved and the unresolved harmonics, consequently, the sum of all these harmonic inputs is also phase-locked reflecting all these contributions. Figure 3.5(c) illustrates the composite time-waveforms of the activation in each of the pitch-neurons (y-axis), all driven by a 200 Hz harmonic complex stimulus. Note that the dominant phase-locking at each pitch-neuron reflects the stimulus 200 Hz fundamental (5 ms period) and its harmonics that match best each template. For example, the 100 Hz pitch-neuron is phase-locked to 200 Hz, reflecting the strong contributions

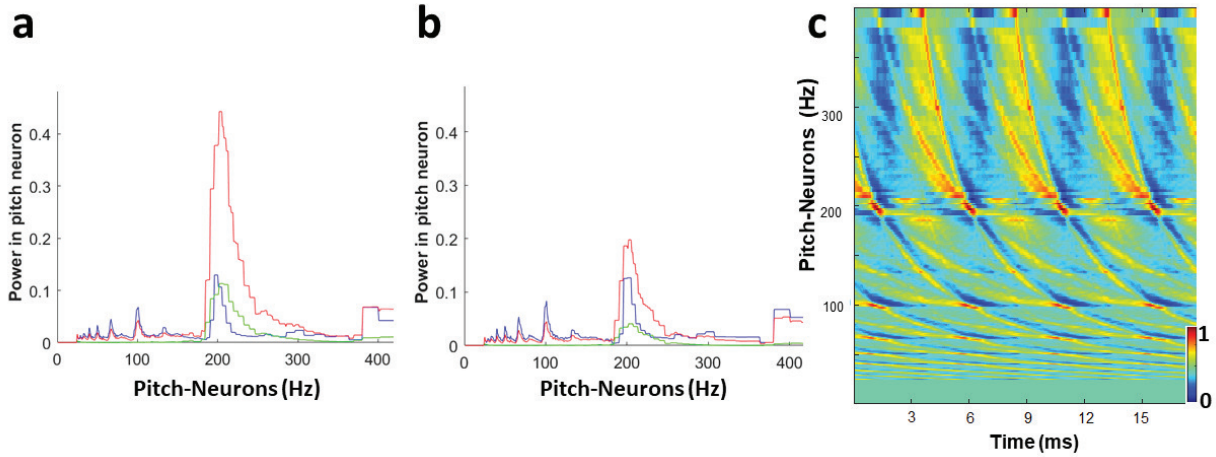


Figure 2.5: Contributions of the resolved and unresolved harmonics to the outputs of pitch-neurons. (a) Responses of an array of pitch-neurons with CF's (or fundamental pitches) ranging up to 400 Hz. The input stimulus is the 15 component harmonic series of 200 Hz. The resulting activation pattern across the array is computed separately for the resolved (blue) using the spectral templates, and unresolved (green) inputs using the dendritic bandpass filters; the combined activation is also shown (red). The pattern here indicates that the maximally activated pitch-neuron is at 200 Hz. (b) Pitch-neuron array outputs for random phase harmonics. The resolved peak height is unaffected (blue trace), while the peak of the unresolved outputs (green traces) is reduced reflecting its decreased saliency. (c) Output temporal waveforms generated in the same array of pitch-neurons whose average outputs are shown in (a,b) above. Color bars next to each spectrogram indicate the scale of the response in arbitrary units.

of the matched harmonics (200, 400,... Hz). Interestingly, scrambling the harmonic phases would change the waveforms even from the resolved inputs, although the location of the peak activations (Figs. 2.5a and 2.4b) remains unchanged.

2.5.1 Physiological responses of pitch-neurons

What is the physiological evidence for the existence of such neurons early in the auditory system? As discussed earlier [79], the structure of the postulated pitch-neurons must exhibit dendritic inputs from a wide region across the tonotopic axis, so as to integrate the harmonics of the CF. Such neurons must exhibit broad tuning with multiple harmonically related peaks and are likely also to have high-thresholds for single tones that decrease as more inputs (harmonic tones) are added. The responses of a pitch-neuron are expected to be phase-locked to the composite of its inputs, but most importantly to its CF. Consequently, they are likely to exhibit well-timed onset responses when the inputs are at zero relative-phase. Finally, since phase-locking is essential for the formation of the input synapses, these neurons must have low frequency CFs ($<$2 kHz) allowing for at least one (phase-locked) harmonic to contribute to the maximum pitch.

In the physiological search and testing of these cells' responses, it is usually very difficult to attain views that match the patterns of Fig. 2.5(c) because it requires recordings from more than a few pitch-neurons. Instead, experiments typically record from a single neuron, and change the parameters of the stimulus to explore the unit's sensitivity. Thus, if one is to sweep the fundamental frequency of the harmonic complex, the pattern of responses of the hypothetical pitch-neuron might resemble that in Figure 2.6, both for the average activation (left panels) and the phase-locked responses (right panels). Figure 3.6(a) displays the responses of a pitch-neuron

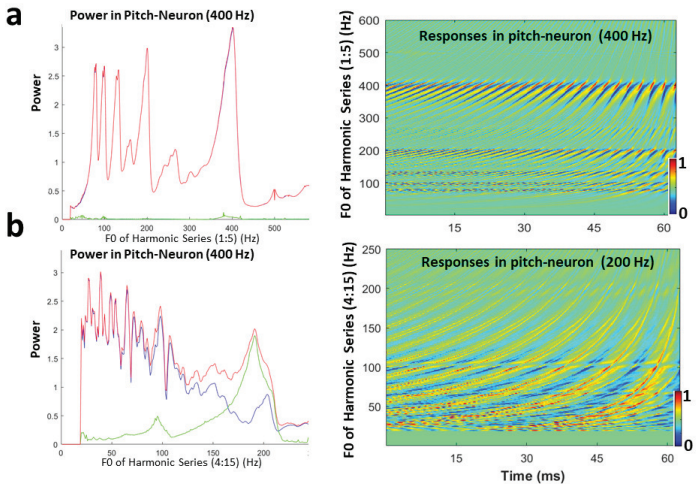


Figure 2.6: Physiological responses of single pitch-neurons tuned at different CFs. (a) Sweeping a harmonic complex (1:5 harmonics) over a range of fundamental frequencies ($50 < F0 < 600$ Hz). (Left panel) The pattern of responses in a single pitch-neuron (CF=400 Hz) is strongly responsive at pitches that are at or multiples of its CF. (Right panel) The waveforms are phase-locked to the CF and occur whenever the harmonics lie within its CF tuning. (b) Phase-locked responses of a pitch-neuron at CF=200 to a partially resolved complex (4th-15th harmonics). Note the unresolved responses (green trace) are weak when the F0 is below the CF of the cell (200 Hz). At those low F0's, the resolved harmonics contribute but the patterns of response and phase-locking are a complicated mix of the fine structure and envelope modulations. Color bars next to each spectrogram indicate the scale of the response in arbitrary units.

(CF = 400 Hz) to a simple five-harmonic complex as its fundamental frequency (F0) is swept from 50 to 600 Hz. As expected, it is excited whenever one of the five harmonics falls within its tuning, hence producing peak responses when the F0 = 80, 100, 133, 200, 400 Hz. Note that the response waveforms (right panel) are all phase-locked to 400 Hz because the neuron is always excited by whichever harmonic lies near its CF (400 Hz) and its multiples. This simple response pattern becomes far more complex if the harmonic series contains a substantial number of unresolved components. This is demonstrated in Figure 3.6(b) for a 200 Hz pitch-neuron driven by the 4th–15th harmonics of a sweeping complex. The response here reflects both the fine structure of the resolved harmonics, as well as the unresolved modulations. We emphasize, however, that this "physiological view" of one pitch-neuron driven by a sweeping F0 (Fig. 2.6(b)) is not equivalent to the view one obtains from the activation pattern across the array of all pitch-neurons as in Fig. 2.5. Consequently, the peaks in Fig. 2.6b patterns are not necessarily indicative of the perceived pitch since the maximum response may occur at lower F0's than the CF of the pitch-neuron. For instance, in Fig. 2.6b the pitch-neuron (CF = 200 Hz) is strongly driven by the F0 = 50 Hz complex because its harmonics happen to coincide with the CF. Nevertheless, this F0 complex would activate the pitch-neuron at 50 Hz far more strongly, and hence if one is to view the activation induced by this complex (F0 = 50 Hz) across all pitch-neurons, one would see a clear peak at CF = 50 Hz.

A different set of physiological experiments in search of the elusive pitch-neurons are those that demonstrated tuning to specific amplitude modulation (AM) rates, mimicking the dendritic filter banks described earlier (Fig. 2.3). These tuned responses, however, have been most commonly described in the inferior colliculus [94, 95] and have not been systematically tested with harmonic tones to demonstrate their relationship to the harmonic templates. Furthermore, the

CFs of these neurons are often quite high (>3–4 kHz), and their bandpass tuning is broad (Q = 1–2), suggesting that they may at best serve to estimate the unresolved-pitch. However, the modulation tuning range described (>1–2 kHz) is not commensurate with the observed psychoacoustic limits of about 500–700 Hz for the unresolved pitch [69]. These conflicting data, together with the absence of any information as to how this modulation tuning arises, or its relevance to other pitch properties, cast substantial uncertainty on this physiological evidence.

Finally, it is worth mentioning that maps and responses sensitive to the pitch of the resolved harmonics have been reported in monkey and human primary auditory cortex [32,81,104]. These findings, however, do not shed light on the mechanisms that give rise to them, because the resolved-pitch percept must arise earlier in the auditory pathway where phase-locking to the harmonic tones is still present, e.g. somewhere before the inferior colliculus.

2.5.2 Pitch of binaural stimuli

A harmonic complex is generally heard at the same pitch when presented monaurally or binaurally. However, it has been discovered that the two percepts diverge if the stimuli presented to the two ears are different. Two classic psychoacoustic studies are discussed here. The first is the Huggins pitch illusion [105,106] in which two spectrally broad noise stimuli are identical except for interaural π phase-shifts introduced at one or multiple very narrow frequency bands (1/6 octaves). The two stimuli are heard as identical noises monaurally; however, the binaural percept contains additional faint tones at the locations of the interaural phase-shifts. Interestingly, when resolved and harmonically spaced, these tones behave just like harmonic complexes, evoking a pitch percept, even if the fundamental of the series is missing. Since such interaural phase-shifts

are usually extracted through binaural convergence [107, 108] in the nuclei of the superior olivary complex (SOC), it has then been commonly concluded that the harmonic templates must be located after or more central to the point of convergence (e.g. at or beyond the SOC complex). A similar conclusion has been drawn from a second set of experiments in which a harmonic complex is broken up into two sets of components. When presented dichotically, the pitch percept elicited is that of the original complete set. For example, if odd and even harmonics of a 200 Hz fundamental are presented separately to the two ears [66], the monaural pitches perceived reflect the specific complexes presented to each ear (e.g. 400 Hz in the even-ear, and approximately a 200 Hz pitch in the odd-ear). However, when presented simultaneously, the harmonics are perceived at the pitch of the original complete 200 Hz complex. This reliable finding has also been cited as evidence that the pitch templates used to estimate the pitch must occur at, or subsequent to the binaural convergence.

However, these conclusions are unwarranted as we demonstrate in Fig. 2.7. We assume that the spectro-temporal templates are located early along the monaural pathway, prior to the SOC. A key observation to make about the responses of the model pitch-neurons is that they are themselves phase-locked to the combined waveform of all of their (resolved and unresolved) inputs, as was demonstrated in Figs.2.5 and 2.6. So, we consider first the Huggins pitch illusion elicited by inter-aural phase-shifts at five harmonic frequencies, e.g. 2nd–6th harmonics of the fundamental 125 Hz (250, 375, 500, 625, 750 Hz). These become evident if we simply subtract the auditory spectrograms of the noise in the two ears (Fig. 2.7a), as described by the Equalization Cancellation Model of binaural masking release [109]. Obviously applying this pattern to the templates would estimate it correctly as perceived. However, because of approximate linearity and preservation of the phase-locking, this sequence of operations (subtraction followed by the

template inner product) can be readily switched while maintaining the same outputs, i.e.⟨ L – R,T ⟩ = ⟨ L,T⟩ -⟨R,T⟩. The subtraction then can be done after the templates, for example in the SOC, or via the inhibitory cross-projections in the LLN or IC [110, 111]. We emphasize again that all this requires that adequate phase-locking is still preserved at the outputs of the pitch-neurons so as to represent the phase-shifts between the binaural noise stimuli. The schematics in Fig. 2.7b illustrate two possible versions of these computations, and we provide in Sec. 2.3 the mathematical formulations implemented to compute the results of the example in Fig. 2.7a. The upper panel illustrates the difference spectrograms between the two ears, and the pitch-neurons activation (lower panel) induced by this spectrogram, showing a large peak at the missing fundamental pitch of 125 Hz (red arrow), its octave (250 Hz), and other related pitch frequencies (black arrows).

In Fig. 2.7c, we compute and interpret the outputs due to the binaural odd-even harmonic stimuli described earlier. The simplest (if not quite accurate) way to intuit the results is through the linear operations mentioned earlier. Thus, when the monaural patterns are applied separately to the templates, and the outputs added, we get the expected percept described by the experiments and illustrated in Fig. 2.7c. Again, this "addition" can occur in the MSO, or even in the identical neural circuits postulated in Fig. 2.7b. In fact, the outputs in Fig. 2.7c are all computed using the identical formulations used for the Huggins pitch examples and provided in Sec. 2.3.

2.6 Discussion

We have described a biologically inspired model for how pitch percepts can be computed via harmonic templates which utilize both temporal and spectral structure of harmonic sound

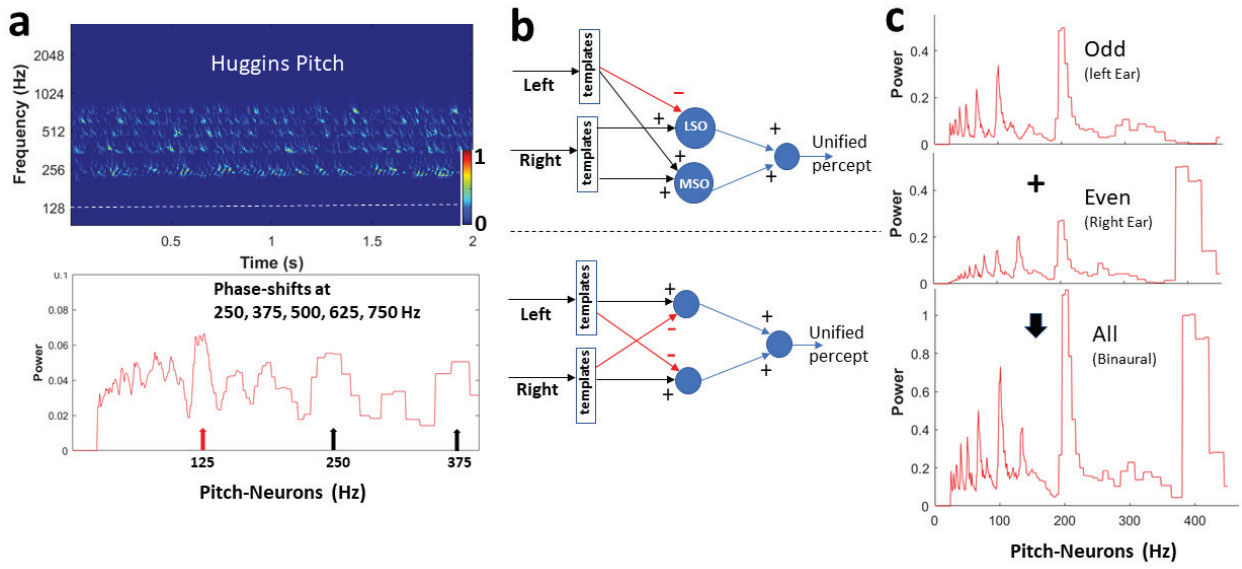


Figure 2.7: Binaural pitch phenomena do not require post-binaural pitch templates. (a) Huggins percept pitch is evoked by highly localized inter-aural phase-shifts, which produce a percept of a tone. These can also be combined with other shifts at harmonic intervals to evoke resolved-pitch percepts and the missing fundamental. (Bottom Panel) Activation in an array of pitch-neurons indicates the percept of the missing fundamental (red arrow) by the Huggins tones (black arrows) at 2nd - 6th harmonics of 125 Hz. (b) Schematic of the computations implemented to generate the binaural percepts of the Huggins pitch and the dichotic odd-even harmonics in next panel. (c) Dichotic asymmetric harmonic complexes. Arrays of pitch-neurons are activated by different monaural harmonics (odd I& even of 1:10 harmonics of F0 = 200 Hz, presented in separate ears). Each combination evokes the expected percepts (peaks in the pattern for Odd = 200 Hz; and for Even = 400 Hz) when presented separately (top and middle panels). When presented simultaneously (bottom), the percept is that of the full complex (200 Hz). Color bars next to each spectrogram indicate the scale of the response in arbitrary units.

stimuli. The model explains how consistent pitches are estimated from resolved and unresolved harmonics, and how they are subsequently fused to give a unified percept. The postulated pitch-selective template neurons receive synaptic inputs from all cochlear regions. Specifically, they receive cochlear inputs conveying the low-order resolved harmonics, forming the classic harmonic templates. They additionally receive the temporal modulations due to the interactions among the higher-order unresolved harmonics, which are subsequently filtered by the dendritic trees of the target neurons, each according to its pitch-selectivity. We have also demonstrated how during development, each pitch-neuron can form all the necessary harmonic-template synapses, and also "tune" its dendritic filter bank through "Hebbian" modification of its synaptic weights. The process can be driven entirely by white noise, with no need for specific harmonic exemplars, and guided solely by the basic tonotopic input into each pitch-neuron.

Pitch estimates generated by the model neurons through such spectro-temporal templates reproduce the well-known properties of the pitch of resolved harmonics, including the percept of the missing fundamental, and its insensitivity to the harmonic phases. Simulations also show that the harmonic templates need not be situated at or post the binaural structures as has been assumed to explain a variety of dichotic binaural percepts [112]. Instead, we show that monaural templates can readily reproduce the binaural percepts if their phase-locked outputs are combined (summed or subtracted) afterwards.

However, the key new contribution of this spectro-temporal template is its integration of unresolved (temporal) pitch into the classic harmonic (spectral) templates via the tuned dendritic filter banks. The existence of these dendritic resonances is almost unavoidable by the very nature and construction of dendritic trees. The resonances modeled by Laudanski et al. [91] do not rely on complex precise processes in any way. Just the opposite, they emerge out of basic prop-

erties of neuronal membranes and branching patterns as well as basic ionic channel properties. Loosely speaking, the combination of these factors inevitably gives rise to resonances that are fairly broadly tuned and that can account for the decoding of the unresolved pitch as described in the upper branches of the template model. Unfortunately, most research on this topic in dendritic resonances has not looked for or exploited such high frequency resonances. Instead, it has focused on dendritic processing in the classic range of cortical rhythms, which are in the range $<$10's Hz. Therefore, the interest in resonances like these is likely to be the exclusive domain of auditory pitch processing! Hence, there is little mention or experimentation to explore this topic beyond what was reported and cited in Laudanski et al. [91].

Finally, the proposed spectro-temporal template elucidates how dendritic filtering imposes a low limit on the resolution and range of the unresolved pitch percept. It also clarifies why temporal modulations at a specific rate are perceived at a pitch equal to that of a tone, or of a harmonic complex with a fundamental, at the same frequency. It should be noted that all these inputs onto the spectro-temporal template or "pitch-neuron" are ultimately unified by their harmonic relationship to the CF of the somatic input of the pitch-neuron (Fig. 2.4b) because it is this input that determines which resolved harmonics succeed in synapsing onto the cell [79], and which resonant dendritic synapses survive (Fig. 2.4). However, as we discuss next, our implementation of this unified spectro-temporal template in terms of a "pitch-neuron" model (Fig. 2.4b; right panel) is purely speculative. It is indeed entirely possible that other more imaginative and distributed neural circuits could carry out these computations [79].

2.6.1 Physiological tests of the spectro-temporal templates

The site at which pitch is first computed along the auditory pathway remains one of the fundamental mysteries of auditory neuroscience. It is remarkable that after many decades of neurophysiological experimentation, and with so much detail known about the psychoacoustics of pitch, we still do not know which structures in the early auditory pathway are responsible for the extraction of this crucial percept. Part of the difficulty stems from the myriad variety of temporal and spectral cues available at the cochlear outputs and at subsequent stages which do not decisively exclude numerous contrasting ideas. Another problem is the complex morphology, physiology, and anatomical structure of the cochlear nucleus and following midbrain nuclei, which contain many interconnected subdivisions yielding a bewildering set of pathways up to the inferior colliculus. Finally, the perception of pitch in humans is apparently far more developed than in most mammalian species used in physiological studies of the early auditory system [17, 102, 113]. It is therefore quite possible that the search for the neural substrates of pitch has been frustrated by their absence in these animals!

Nevertheless, assuming that these pitch substrates exist, how can this model guide a new search that benefits from and complements the extensive knowledge that has already been gathered over decades? To start with, there are two clear key requirements that dictate where the templates are likely to form. The first is temporal; precise phase-locking is necessary in the formation of the pitch templates. And for many well-understood reasons, phase-locking begins to diminish after the cochlear nucleus (CN), and it is substantially worse by the time we arrive at the inferior colliculus (IC). Consequently, it is likely that the templates form in the CN or just after, but certainly not beyond the IC. The second key requirement is spectral; broad convergence

of cochlear channels is an inevitable ingredient in the formation of pitch templates. The cells that would respond as pitch-neurons of the model must therefore integrate harmonically related inputs and be phase-locked to the aggregate of these inputs.

There are only a few cell-types in the early auditory pathway, and especially in the CN, that have these properties, and that have been postulated to serve as the harmonic templates. They include the various Onset cells (one of which is also known as the Octopus cell [110, 114–117]) of the Postero-ventral CN (PVCN), which respond robustly to multiple broadly-tuned synchronous inputs and show strong phase-locking to click trains at least up to 2 kHz [118, 119]. The morphology of the octopus cells is also consistent with the requirements of the pitch templates; that it has inputs across the frequency spectrum, and have a cell body located in the lower range of tonotopic organization, and an extensive branching network of dendrites radiating perpendicular to the tonotopic access, thereby accessing the full range of frequency inputs [114, 115]. However Octopus cell projections are excitatory to the ipsilateral ventral nucleus of the lateral lemniscus (VNLL), where their targets provide fast, inhibitory inputs to the contralateral inferior colliculus. These projections do comport with a functional role of onset detection and reset of neuronal states, but not necessarily for periodicity estimation. Furthermore, it is unclear if the somatic and dendritic inputs to these cells from the auditory nerve are spectrally sharp enough as those illustrated by the auditory spectrograms of figs. 2.1 to 2.3 and 2.4b. In our simulations, the auditory spectrograms are sharpened and rendered level-invariant through the action of a LIN [82,83] (as explained in more detail in Sec. 2.3). The LIN would have to precede the Octopus cells so as to provide the sharpened spectrograms, and no such substrate is known to exist.

Another possible candidate are the neurons of the VNLL, which form extensive and spatially-

organized connectivity that has been described as the double-helix of the auditory pathway [120, 121]. The VNLL's primary inputs are the octopus cells in the PVCN and the T-stellate cells of the antero-ventral CN (AVCN). These T-stellate neurons are particularly interesting in that they have been postulated to perform the LIN necessary to extract from cochlear phase-locked responses a robust, sharp spectral (Q = 10–12), and yet still phase-locked representation regardless of stimulus levels, similar to that generated by the cochlear model in this report and shown in all auditory spectrograms in figs. 2.1 to 2.4, (see also Sec. 2.3) [82, 122]. They have been described as encoding a spectral representation of sound [121]. VNLL neurons receiving excitatory inputs from T-stellate and octopus cells exhibit multiple morphologies and a large variety of physiological response types whose function remain largely a mystery. Nevertheless, the anatomical organization [123, 124], the biophysical properties [125], and the few physiological measurements made with harmonically complex stimuli [126, 127], have inspired proposed schemes of pitch encoding that may be relevant to the model presented [120].

In summary, it seems that the search for the pitch-neurons in the early auditory pathway may well be productive if it is guided by the findings already available and the general constraints described above. However, for a variety of reasons, there are very few studies of VNLL and CN responses that have directly and simply tested whether the neurons are in fact tuned to harmonic complexes. There are, of course, many difficulties and confounds that can hinder such an investigation and search, including cochlear and neural nonlinearities that render many analysis methods (e.g. Fourier analysis and interval histograms) extremely confusing and distracting. Nevertheless, it is possible to disambiguate many of these measurements and their interpretations with simulations similar to those in Fig. 2.6.

2.6.2 Psychoacoustic challenges of the spectro-temporal templates

As discussed above, much of what we know about the psychoacoustics of pitch perception can be fully accounted for separately by the spectral templates for the resolved pitch, and the temporal filter banks for the unresolved pitch. However, recent investigations have highlighted a challenge to the well-accepted notion that phase-locking is critical for explaining pitch. These concern the finding that pitch perception is possible with resolved harmonics of frequencies well-above the phase-locking range in humans (8–10 kHz). For example, the ability to hear a 2 kHz pitch as the missing fundamental of a 4th–6th harmonic complex [5, 6]. In our model, the (spontaneous) formation of the spectral templates, and specifically of the synapses of the harmonic inputs, requires that the inputs be phase-locked so as to correlate with the CF of the pitch-neuron. Of course, once the inputs become connected, phase-locking is not essential to convey the harmonic energy.

The simplest way to circumvent the phase-locking limitation above is to postulate that some pitch templates form by repeated exposure to the harmonic complex itself, and not simply by the generalized noise we described before in Ref. [79]. Thus, it is commonly found that various patterns of activation (spectral such as vowels or the ultrasonic harmonics of the echolocating Mustache bats [34]) can in principle be learned from examples. In the same way, exposure to the harmonic complex of 2 kHz can link the template (formed already from the low-order harmonics) to the high-order non-phase-locked components when they are presented together. This would explain how these high harmonics of the complex induce the same pitch of the missing-fundamental, and why the resolution of the pitch percept is typical of other resolved pitches.

2.7 Acknowledgments

This work was supported in part by NIH R01DC005779, Advanced ERC grant (Neume), NIDCD training grant No. DC00046, and NSF 1632976.

Chapter 3: Methods of Electrophysiology

Chapters chapters 4 and 6 utilize single unit neuronal spiking data recorded *in vivo* from unanesthetized adult female ferrets (*Mustela putorius furo*). The figure-ground study in chapter 4 includes data from 3 animals and the selective attention study (Chapter 6) includes data from 5 animals, one of which was also used in chapter 4. The total number of animals used among the three studies was 7.

3.1 Training and behavioral tasks

Animals were trained on a go/no-go appetitive selective attention streaming task (figure figs. 5.1 and 5.2), the data from which is presented in Chapter 4. Three animals were trained on the intensity detection version and two animals were trained on the gap detection version. In both variants, animals were trained to withhold licking from a water spout during presentation of a variable number of reference sounds (1-5 s). Because this is a very difficult behavior to learn, we allow a small amount of licking during the reference sounds before considering a trial a false alarm. Specifically, the reference sound presentation window is divided into 1-second bins. If during a given trial, the animal licks during less than half of those bins, the trial continues and they are able to obtain a reward with a correct response. If they lick during half or more of the response bins, the trial is immediately terminated and considered a false alarm. Correct responses

occur when animals lick in response to a target sound, triggering water to flow from the spout for a 2 s reward window. Lick responses less than 20 ms after the target onset are considered failed trials because the response occurred faster than the animals' plausible reaction time [18].

In the selective attention streaming task, the reference stimulus consists of pairs of alternating pure tones with a cosine ramp of 5ms and a short gap of silence between tones. The length and rates of the tone presentation varied between experiments but all were held constant for each animal after training was completed. A narrowband noise (width ½ octave) burst plays synchronously (coherent) with one of the pure tones. All reference tones and noise bursts are at a constant intensity until the target occurs. After 10-30 alternating burst epochs, the target noise burst occurs and the animal can lick to obtain a reward. In the intensity cue task, the target was a noise burst presented 15dB higher intensity. In the gap detection task, the target was a noise burst with a 100 ms gap centered relative to the onset and offset. During electrophysiology recordings, the identical stimulus task set is presented before and after behavior, while the animal passively listens.

3.2 Surgery

Behavioral animals were trained in a freely-moving behavioral rig until they reached a threshold level of task performance. To facilitate stable recording conditions, animals were surgically implanted with a stainless steel headpost attached to the midline suture of the parietal bones. The implant material (Charisma or Zimmer bone cement) helps to create a sterile field for maintaining a chronic craniotomy.

One day prior to surgery, animals are given antibiotics (Cefazolin) to facilitate recovery

and prevent infection. Anesthesia was induced with dexmedetomidine and ketamine and was maintained with 2% isoflurane. Respiratory rate, ECG, heart rate, spO2, and body temperature were monitored from induction to when the animal woke up, and emergency drugs were given as needed to maintain vitals in a safe range. Using sterile procedures, the skull was exposed and temporal muscles removed to create a surface for the implant material. A stainless steel headpost was affixed to the skull and titanium screws were embedded to form a mechanical surface to attach the implant. Dental cement-like material was applied to the top of the skull, with fewer layers over areas of skull covering auditory and frontal cortices. The location of the temporal lobe and AC is estimated and measured using anatomical markers. Animals were in recovery for 4 weeks post-surgery, with additional antibiotics and analgesics administered as needed. After recovery, animals were habituated to head-fixation in a custom lucite cylindrical holder for increasing durations over a period of 2 weeks. Then, they are retrained in the head-fixed experimental rig until achieving threshold performance, at which time recording can begin.

Before recording, a microcraniotomy (1-3mm diameter) was created for access to cortical recording sites by first drilling through the layer of implant material covering the skull directly above the auditory cortex. Cavities in the implant material are kept plugged and covered between experiments using silicone ear impression material and were cleaned and treated with topical antiseptic drugs (betadine solution) at least once per week and after every recording session. The skin surrounding the implant was cleaned 3 times per week with saline, dilute betadine solution, and sulfadiazine ointment. After the craniotomy the animal is allowed to rest for at least one day to allow the brain to recover from possible damage during drilling.

3.3 Structure of recording sessions

Recording sessions typically had a duration of 4-8 hours. All experiments were conducted in a double-walled sound attenuation chamber (IAC). The animal was secured in a custom lucite hold and head fixed in place by securing the headpost to a steel bar affixed to the holder. The craniotomy surface was cleaned and four tungsten microelectrodes with high impedance (4-7 mOhms) are used to record extracellular potentials in the brain tissue. Electrodes were independently advanced at 30 degrees from vertical, approximately perpendicular to the cortical surface, into the brain using an EPS driver system until spontaneous spiking activity was seen in all or most channels.

Extracellular spiking data was acquired at 1-3 penetration depths per recording session. Passive tuning properties were assessed using responses to pure tones, temporally orthogonal ripples combinations (TORCs), and narrowband noise bursts. After measuring passive tuning properties, the experimental stimuli were presented, including passive presentation in all sessions and active behavioral conditions for some. After the last passive presentation for one recording depth, electrodes were advanced 200 μm in order to find a new set of neurons. After recording, electrodes were retracted from the brain and the craniotomy was cleaned and resealed with ear impression silicone. Stimuli, data acquisition, and analysis are done using custom MATLAB software, and data acquisition also involved AlphaLab software from Alpha Omega. High impedance tungsten electrodes (FHC) were used for extracellular recording, arranged in a 2x2 square array and separated by 0.5 mm.

Data acquisition hardware from Alpha Omega recorded electrical potentials at 22 kHz and amplified x15000. Raw data was filtered and single units were isolated by hand using a combina-

tion of principal component analysis, K-means clustering, and template matching using custom MATLAB software. MATLAB code implementing online multiunit analysis was used to calculated tuning curves and spectro-temporal receptive fields (STRFs) for each recording site during the session. Recording location in primary auditory cortex is determined using anatomical markers and comparing responses to known response patterns in primary and secondary cortical areas [49], [50]. Each neuron or multiunit signal is characterized by a best frequency, which is used to choose the stimulus parameters for a recording session.

3.3.1 Localization of A1

The location of A1 was first determined by relative distance to cranial landmarks during implant surgery. The first craniotomy for each animal was placed according to these landmarks and location in A1 was determined during the first few recording sessions by analyzing the tuning properties of the recorded cells using pure tone tone pips and TORCs to generate tuning curves and STRFs. Neurons in A1 typically exhibit sharp tuning to pure tones and single-peak STRFs with short latencies. Subsequent recordings allowed us to determine recording location in A1 based on its characteristic tonotopic organization with high frequencies most most dorsal and low frequencies ventral.

All sound stimuli were presented at 65dB sound pressure level (SPL), with the exception of the pure tones tones for generating tuning curves, which were presented at 25-65 dB. Sound stimuli were generated in MATLAB at 40 kHz sampling rate using custom functions and presented using National Instruments A/D hardware (PCI-6052E) and a free field speaker located 1 m away in front of the animal's head.

3.4 Data Analysis

Tuning properties were assessed using spiking responses to 100 ms pure tones spanning 4-8 octaves, with 11 tones per octave, presented at a range of 40 dB in 10 dB steps. The best frequency (BF) was determined based on the mean of a Gaussian curve fit to the average firing rate in the 100 ms window after tone onset.

Baseline STRFs were computed using responses to TORCs presented at a frequency range overlapping the cells' BFs. STRFs were obtained using reverse correlation [103] with the spectrograms of the TORC stimuli and multiunit or single unit responses. Positive values (red areas) of the STRFs correspond to spectrotemporal regions that caused firing rate above baseline or excitation, while negative (blue areas) correspond to decreased firing or suppression.

Chapter 4: Stochastic Figure-ground segregation

4.1 Introduction

Most studies of non-speech streaming processes rely on subjects' reports of perception and use highly sparse acoustic scenes. Obtaining objective measures of streaming require careful experimental design [128,129]. Micheyl et al. [128] assess streaming by using tasks where stream formation would either facilitate or impede performance. Briefly, they used two-tone repeating triplet (ABA) stimuli and required subjects to report a difference in the timing either within-triplet (impaired by streaming) or between-triplets (facilitated by streaming). Using this clever task design, the authors showed that in an alternating tone, streaming-facilitated task, increasing Δf degraded performance. Similarly in a streaming-hampered task, increasing Δf made the task easier. On average, these objective performance-based measures correlated with subjects' reports of single or multi-stream perception [128].

Coherent moving dot stimuli were first used by Newsome and Paré in 1988 to investigate neural correlates of coherent movement detection [130]. In recent years, Teki et al. 2013 adapted this stimulus for investigations of auditory temporal coherence, calling it a stochastic figure-ground (SFG) [131]. In their stimulus, multi-tone chords are presented periodically, at first with all random frequencies and later with a subset repeated in consecutive chords. This stimulus produces a perceptual pop-out, the figure, among non-attending human listeners, indicating that

some pre-attentive mechanism is at play. A modified chord SFG stimulus has also been used successfully with ramped and locally roving figure frequencies [132, 133], both of which are more naturalistic sounds and approach speech-like streaming. Monkeys trained on the chord SFG task showed elevated AC responses during behavior even on miss trials, indicating that AC neurons respond to coherence even without (successful) attention [134]. Shinn-Cunningham 2008 asserted that "attention operates on objects," however it may be that perceptual objects only emerge with subconscious attention [135].

We introduce here a modified version of the SFG stimulus, replacing periodic chords with a cloud of tones. Embedded in this tone cloud, a subset of 4-10 tones are presented coherently, with synchronized onsets, and form a perceptual pop-out known as the figure. The number of tones in the figure is called the coherence level (CL) and is generally proportional to the ease of hearing the figure [49, 136]. This tone cloud stimulus is analogous to coarsened noise and is arguably a more naturalistic acoustic background than random chords. It has been used in previous studies [53, 137, 138] to measure tuning in AC neurons and can be generated over a rich parameter space. The SFG stimulus can be used to probe mechanisms of streaming and perception by manipulating acoustical properties such as the frequency range and spectral and temporal density. It mimics the spectral complexity of the cocktail party problem while avoiding potential stimulus-specific processing strategies, such as linguistic information in speech-in-noise comprehension tasks.

A key difference between the tone cloud and chord SFG [139] is that in the latter, tone onset times are random rather than periodic [134]. It has been shown that listeners can detect a single periodic pure tone embedded in a tone cloud [140], indicating that temporal coherence is not necessary to detect the chord SFG figure. Rather, periodicity or presentation rate may be playing

a role in producing a perceptual pop-out. In our experiment, the average tone rate in each channel is constant, so some cross-spectral channel communication is necessary to detect the figure. Thus with random onsets, we hope to avoid engaging periodicity detection mechanisms (see 1.2.2). Previous work has validated the relevance of the SFG to naturalistic listening by demonstrating correlation between performance on SFG tasks and speech-in-noise comprehension [136], underscoring its relevance to human perception and health.

Another key difference between the stimulus used here and those in previous experiments [131, 141] is that the figure tone frequencies are part of the tone cloud prior to the first figure onset. Thus, the beginning of the figure does not result in introducing new frequencies to the stimulus.

4.2 Methods

4.2.1 Electrophysiology

Animal surgery and electrophysiological recording are performed as described in sections 3.2 and 3.3. The data was collected from 3 animals (n = 3).

4.2.2 Stimuli

Cloud of tones: A cloud of tones consists of a group of fixed-length (50 ms) tone pips with 5 ms cosine ramp, which occur at random onset times within a set of discrete frequency channels. A cloud is generated by first selecting an octave range and a fixed number of tones logarithmically spaced per octave. For animals 1 and 2 we used 6 octaves with 6 semitones per octave, centered at 1600 Hz. For animal 3 we used the same parameters as well as two spectrally compressed

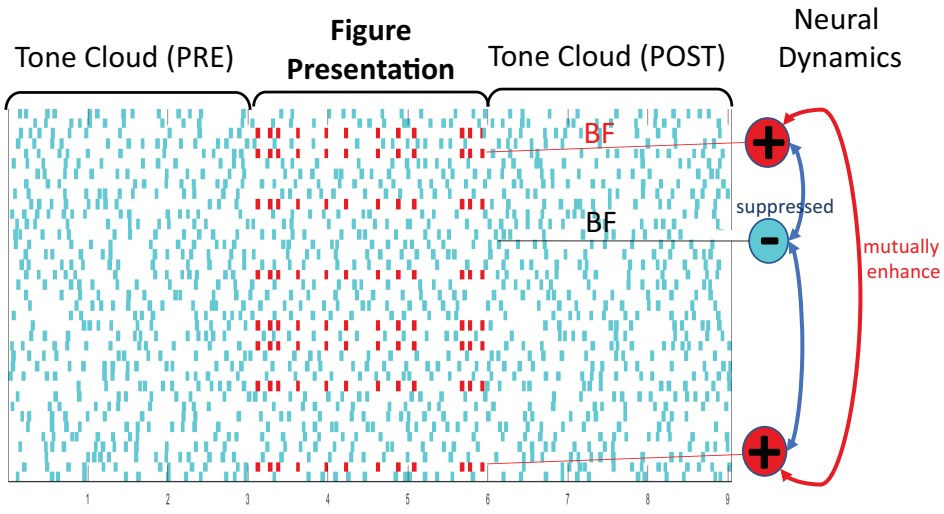


Figure 4.1: Schematic of the SFG stimulus, shown with 3 s portions instead of 5 s for visualization. *right* Predicted neuronal interaction. Neurons tuned to figure tones in red will fire coherently and adaptively and mutually enhance. Meanwhile a neuron tuned to a background tone in blue will be suppressed due to incoherent firing

versions, with 3 octave range and 12 semitones per octave with lowest frequencies 500 and 1500 Hz. Every frequency channel was a mean tone rate of 4 Hz, meaning on average 4 pip onsets occur per second in each channel. Onset times were randomly and uniformly generated for a 2 s window within minimum 50 ms spacing between consecutive onsets, and then are jittered +/- 25 ms to prevent randomly-generated coherence (figure 4.1, first and last third). To obtain a 5 s cloud stimulus, we generated 6 seconds and kept the first 5, eliminating any pips which overlapped the border. *Perceptual Figure:* For the "figure" portion of the stimulus, an initial set of random onset times is generated at an average rate of 4Hz as described above. In order to control the precise timing of the first figure onset and last offset, we require that one onset occurs at t = 0 and one at t = 4.95 s (3 s and 6 s in Fig. 4.1, respectively). Tone pips occur in each of the figure frequency channels at these onset times (figure 4.1, red tones). Then, the remaining channels are filled in

randomly as described above. Importantly, the cloud onsets may not overlap with the figure, in order to keep the final stimulus envelope relatively flat.

Figure frequency channels were chosen by randomly generating a subset of 10 of the tone cloud channels. For the 8 tone figure, a subset of 8 of these 10 channels is randomly chosen, etc. for the other tone number conditions. For ease of interpretation, this procedure was performed one time, and then the same figure tones were used for the remainder of experiments. For the first animal, new random tone clouds were generated for every trial and experiment, however with later animals we pre-generated a set of tone clouds and used these frozen random stimuli for all subsequent experiments.

Each trial consists of 5 s of random cloud, followed by 5 s of cloud with an embedded figure, then 5 s random cloud. We hypothesized that the temporal coherence of the figure tones would cause them to be bound into a stream, and that this effect size would systematically vary with the coherence level in the stimulus, i.e. the number of tones which make up the figure. There are a few different results which would show this: (1). The brain will "clean" the cloud from the cloud+figure portion, i.e. neural responses to non-figure tones will be suppressed during the middle 5 s; (2). The brain will enhance" the figure tones, i.e. neural responses to figure tones will be enhanced during and immediately after the middle 5 s. In each experiment with animal 1, 3 CLs were presented with at least 30 unique stimuli each, for a total of 90 trials. For animals 2 and 3, five unique stimuli were generated for each CL condition and each presented ten times per recording depth, resulting in 50 trials per CL. With animals 2 and 3, at least two CLs were presented in each experiment, in pseudo-randomized trial order.

4.3 Results

The cloud of tones is a naturalistic auditory stimulus akin to coarse noise, which is commonly used to assess tuning. We first validated that our version of this stimulus can re-capitulate response fields of A1 generated by our standard tuning battery, specifically TORC responses (see section 3.3). Using lag values of -10 to 120 ms, the cross covariance of the response and stimulus (auditory spectrogram) was divided by the auto-covariance of the stimulus. Ridge regression was performed with 30 log-spaced values of λ and the results averaged. This procedure was performed for each trial and then averaged, using stimuli and responses to 1-second windows immediately before the figure onset (pre), immediately after the first onset (early), halfway through figure presentation (mid), the last second of figure presentation (late), and immediately following the last figure onset (post). As can be seen in figure 4.2, the tone cloud responses well reconstruct STRFs generated from TORC responses, including the classic checkerboard pattern. Previous studies have indicated that tuning maps generated with tone cloud stimuli can vary systematically with tone cloud density, with more spectrotemporally dense clouds producing sharper tuning [138]. We chose our density parameters based on intuition of human perceptual limits, but it is possible that a denser cloud would produce better tuning and more striking effects (see section 6.1).

An important note is that TORCs contain a continuous range of frequencies spanning a 5-octave range, while the SFG contains only a discrete set of 37 frequencies spanning either 3 or 6 octaves. Thus, excitatory or inhibitory regions which span multiple channels and appear smooth are actually interpolated from a few frequency samples by the spectral blurring inherent in the auditory spectrogram. Horizontally striated regions are visible when this blurring does not fully

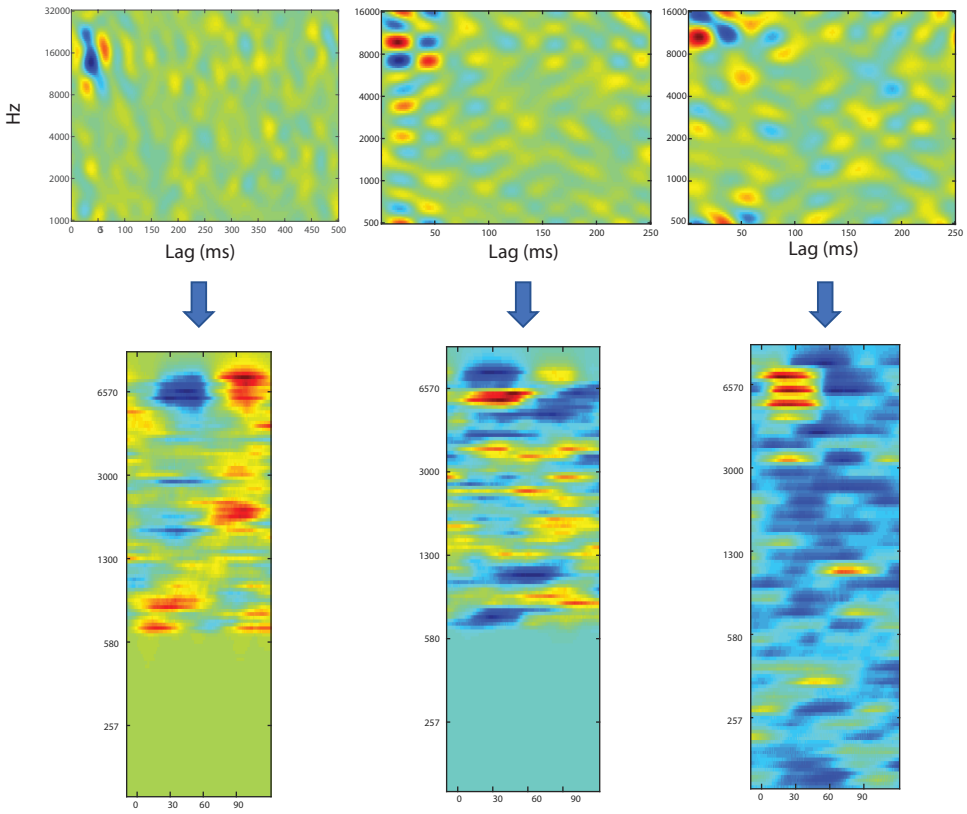


Figure 4.2: Comparison of SFG STRF mapping. For three different units, *top row*: STRFs generated using 3-second TORC responses and *bottom row* STRFs generated using 1-second tone cloud responses. The tone cloud maps capture prominent excitatory and inhibitory regions, with some coarseness due to discrete rather than continuous frequency coverage.

bridge gaps between neighboring tones (see figure 4.2, rightmost plot). 6.1).

Neuronal units demonstrated a few typified responses to SFG stimuli. Units that were tuned to frequencies far from any figure tones had relatively stable response fields throughout the tone cloud and figure stimulus presentation. Some units, including one shown in figs. 4.4 and 4.6 showed strongly suppressed responses to the figure tones while maintaining excitatory responses to their best frequency. Still others as demonstrated in figs. 4.3 and 4.5 were tuned near to a figure tone and showed strong excitation in response to figure tones. While these response types are encouraging as a validation of the stimulus and analysis methods, they do not necessarily

show that a streaming process occurred. One more appropriate evidence is the time-course of response field modulation taking 1s or more to fully adapt (see fig 4.3). This is more indicative of a streaming-like response because it follows the known time course of perceptual streaming.

Because the figure tones are correlated during the figure presentation, it is not possible to disambiguate responses driven by any one of the tones. Therefore, one approach is to examine the response field during the post period to identify any tuning modulation driven by the figure presentation. Previous studies have shown that attention-dependent plasticity can persist for minutes or up to hours [9, 13], so we assume that the one-second window immediately following the figure will retain any modulatory effects.

In order to test for statistical significance of excitatory and inhibitory responses, we generated a null distribution of voxel values using the same procedure as described above for STRFs, but mismatching responses and stimuli. Briefly, for a single trial of spiking data, we generated 4 STRF's using the 4 frozen tone clouds which were not presented to evoke the response spike train. We then down-selected the 128 frequency bins in the auditory spectrogram by removing all channels which did not correspond to the pure tone frequencies in the stimulus. A Wilcoxon rank sum t-test was used to select voxels from the data with significant distributions ($p < 0.05$). Summing over the 37 (frequency bins) by 130 (lags) response map yielded mean correlation values for each of the 5 epochs described above. Subtracting the *pre* epoch from the *post* epoch yields a convenient summary value for the size and direction of change in response - stimulus correlation. Here, we averaged over the response from 20 to 80 ms lag in order to select the strongest response area corresponding to typical A1 tuning.

On a population level, we found that the degree of change from post to pre epochs increased systematically with coherence level. For 4 figure tones, which is perceptually difficult for humans

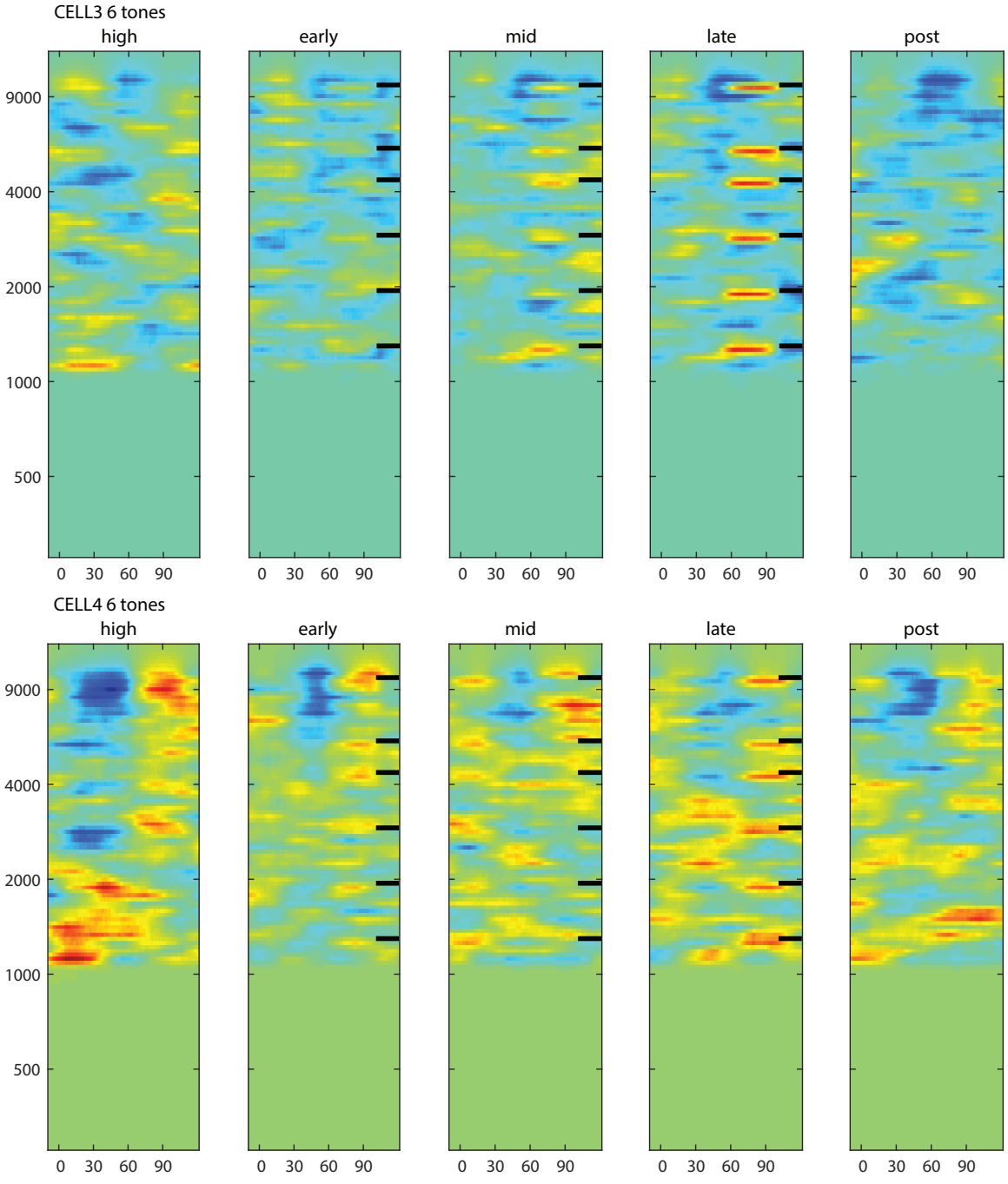


Figure 4.3: Enhanced response modulation evolves over figure presentation. Two example units showing STRFs generated using 1-second windows during the tone cloud (pre), beginning, middle, and end of figure presentation, and immediately after the last figure offset (post). Figure tones are indicated by black bars on the right. For both units, the figure-driven response is initially weak and is strongest in the late period, indicating a 4-5 second buildup of response modulation.

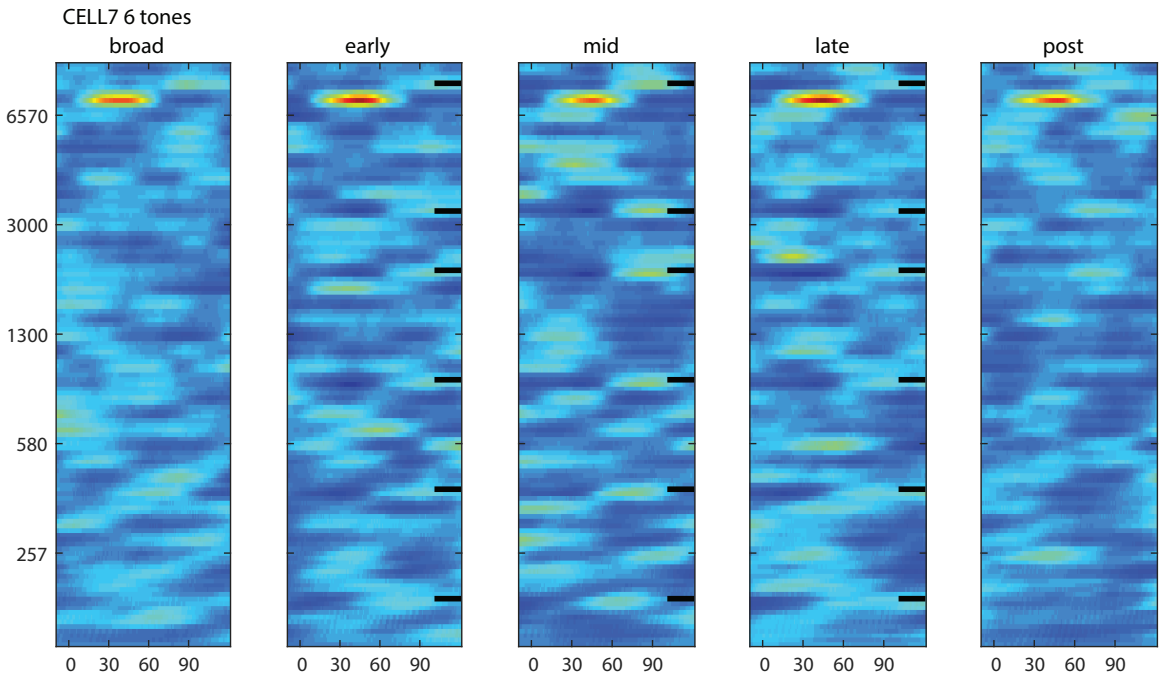


Figure 4.4: An example unit where one figure tone coincides with an inhibitory sideband of the STRF, resulting in inhibitory responses to the figure. Here, response buildup is minimal because the figure tone is well-aligned with a portion of the STRF

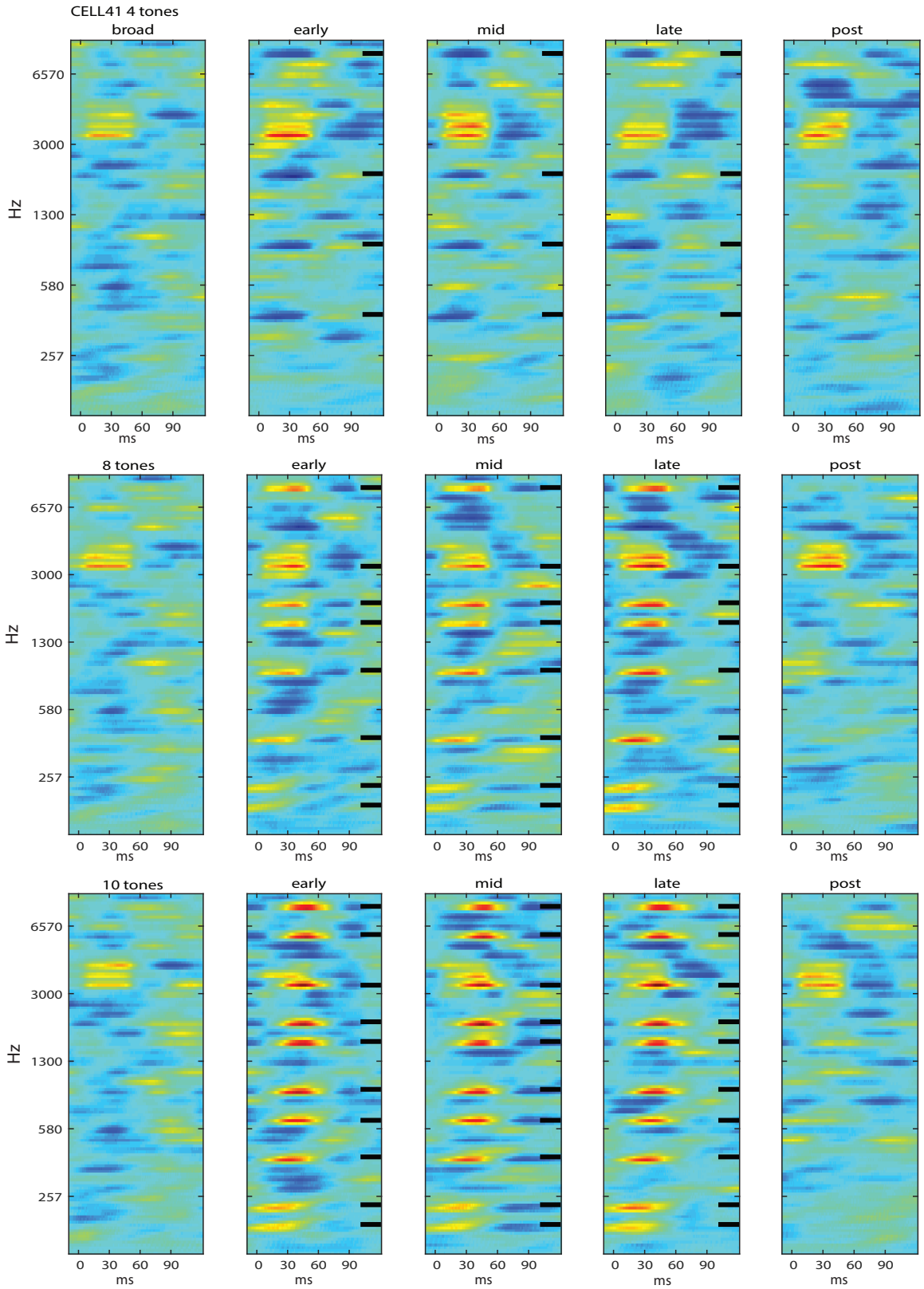


Figure 4.5: Systematic response to increasing CL. All plots from each cell are normalized to the same scale. *Top row* With 4 tones none of the figure tones aligns with the cell's BF and there is little figure-driven response; indeed the unit is suppressed in response to the figure. *Middle row* With 6 tones, one of the figure tones matches the BF and there is a build-up of excitatory response over the 5-second figure presentation window due to the alignment of one figure tone (second from the top, around 3 kHz) with the neuron's BF. *Bottom row* With 10 tones, there is again an alignment with the unit's BF. The unit's excitatory response rapidly saturates and is

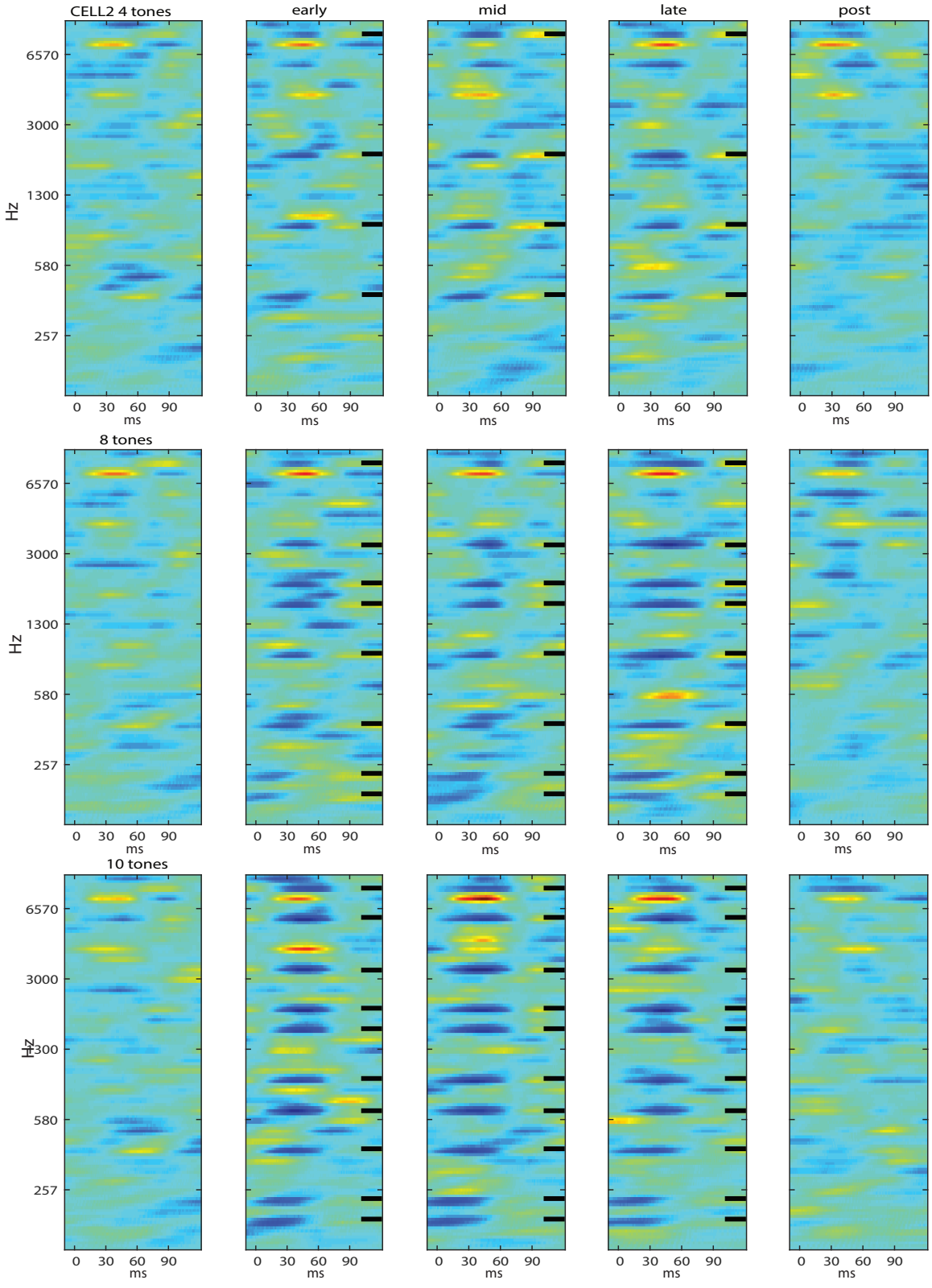


Figure 4.6: Suppressing modulation strength varies systematically with CL. All plots are normalized to the same scale. One example unit responses to 3 different CLs; 4, 8, and 10. In all 3 conditions, one of the figure tones aligns with an inhibitory sideband of the unit's response field. *op row* With 4 tones there is a slow inhibitory or suppression figure-driven response. *middle row* With 6 tones there is a faster and buildup of response and longer suppressed period as seen be the increased length of blue stripes in the *late* response. *bottom row* With 10 tones, the unit's suppressive response rapidly saturates and is maintained throughout the figure presentation window.

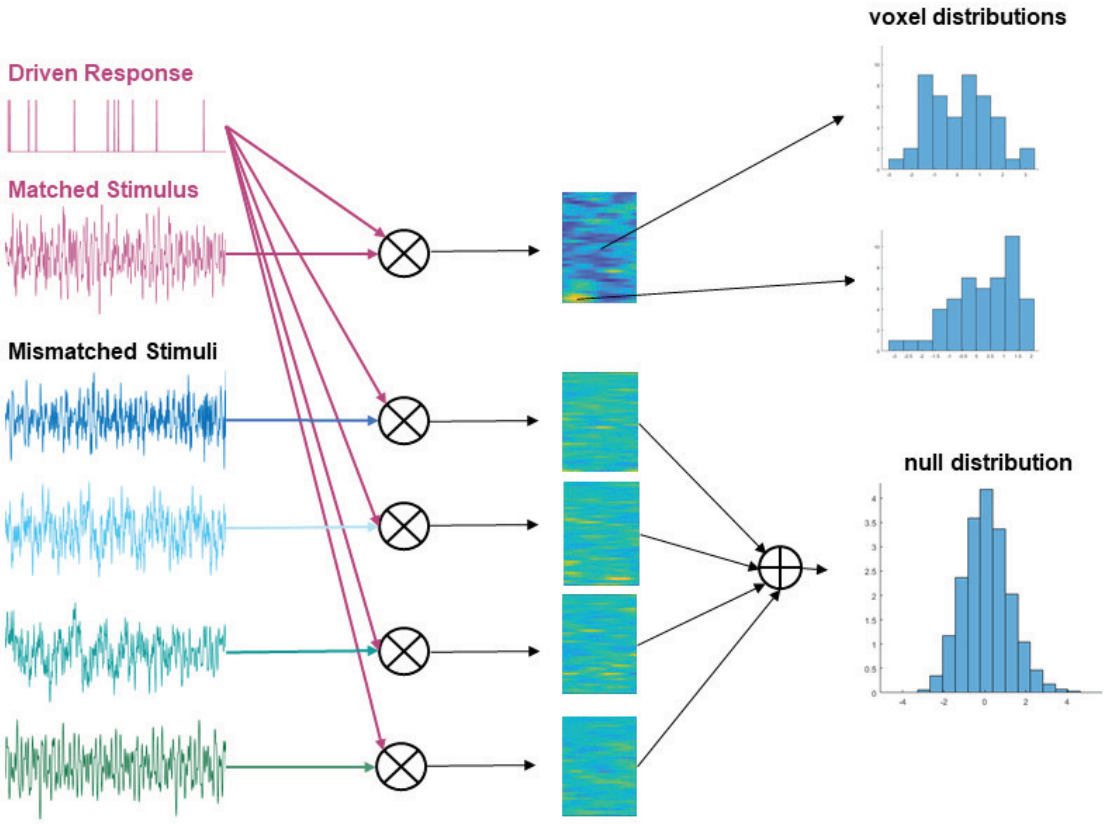
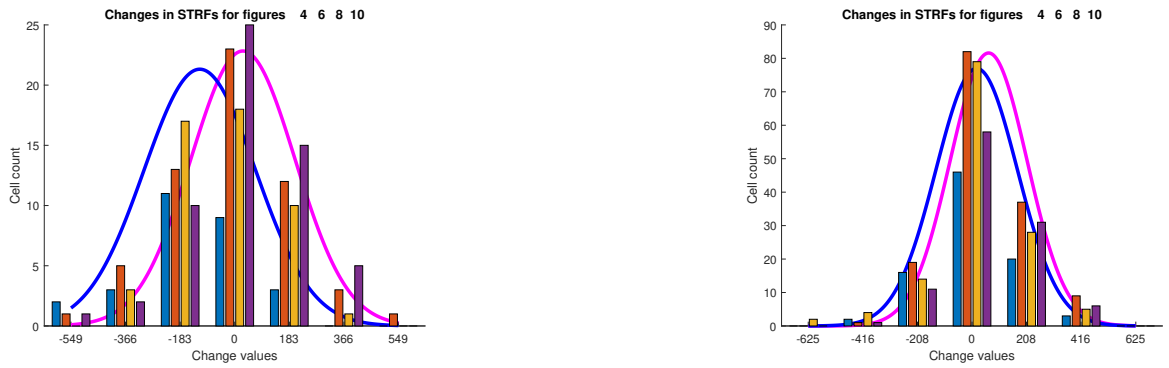


Figure 4.7: STRF selection algorithm. For each trial, STRFs are generated using the correct stimulus spectrogram (displayed here as a waveform for simplicity) in addition to 4, incorrect and mismatched stimuli. The matched stimulus-response STRFs yield voxel distributions corresponding to each frequency – lag pair. The mismatch STRFs are used to generate a null distribution for comparison. In this example unit, an excitatory region in the lower left of the STRF has a right-skewed distribution, while a voxel near the center of the STRF has a flat, centered distribution.



(a) Animal 1

(b) Animal 2

Figure 4.8: STRF changes from pre- to post- figure presentation for two animals. The third animal did not show significant differences between the 4 conditions. Bars from left to right and in order blue, orange, yellow, purple represent responses to 4, 6, 8, and tones. Blue and magenta curves show best fit normal distributions for the 4 and 10 tone distributions (scaled for visibility). These distributions are significantly different at $p < 0.05$.

[136], the post - pre curve shown in figure 4.8 is centered, with a low peak. For 10 tones, which produces a strong perceptual pop-out [136], the curve is right-shifted, which a strong peak in the positive region. These results indicate that with higher coherence levels, the population response was modulated by the figure tones with a net response enhancement that persist after the figure turns off. For 2 of the 3 animals, this effect was strong and statistically robust, however for one animal the effect is unclear. This animal was previously used for the experiment outlined in chapter 4 and may have had age-related hearing loss as well as processing deficits due to extensive recording. For animals 2 and 3, the response changes from 4 to 10 tones were both significant different with $p < 0.005$ (Wilcoxon rank sum test).

The results in figure 4.7 are ambiguous in one respect. Most cells changed in a positive direction, but a large number were in the opposite direction. Why? To explain and predict these opposite changes, we hypothesized that it is due to the tuning of each cell relative to how the figure tones drive it. Thus, when even one the figure tones is aligned with a cell's BF, as is likely the case when there are many tones (8 or 10), the cells are driven well by the figure and the STRF

shows strong excitatory correlations with all the figure tones. The opposite happens with cells that are uncorrelated with the figure tones (or appear to have suppression that is aligned with the figure tones). These cells become gradually suppressed presumably by cells that are excited by the figure tones (as was explained in Figure 4.1. As we explained earlier, this alignment of one of the figure tones and the BF is purely coincidental. And since the tones are correlated among themselves, they all appear to excite or "suppress" the cell similarly in the STRF figure.

To test and exploit this hypothesis, we first estimated the cells' BF by correlating the response during the 5 second *pre epoch* of the tone cloud with each of the tones in each frequency channel. The channel with highest correlation was designated as the BF. We then generated a BF-triggered PSTH of responses during figure presentation by binning and averaging the unit response for ±120 ms surrounding each BF channel onset. We followed the same procedure to generate a separate figure-triggered PSTH, using the onset times of the figure to center the window. We finally compared the two PSTH curves by computing their inner product: the resulting measure is positive if the cell is driven by the figure tones. It is negative if the cell is inhibited by the figure. Finally, our hypothesis is that the post-pre figure changes are enhancements if the correlation measures above are positive, and suppression otherwise. So by splitting the cells into two groups – positively and negatively correlated with the figure. The distributions of the post-pre changes of these two cell groups are illustrated in fig 4.8. Units with values between ± 0.005 were discarded. As can be seen, there is a strong opposite trend in the plasticity of these cells depending on how they are sorted (or driven by the figure) as predicted by fig 4.10.

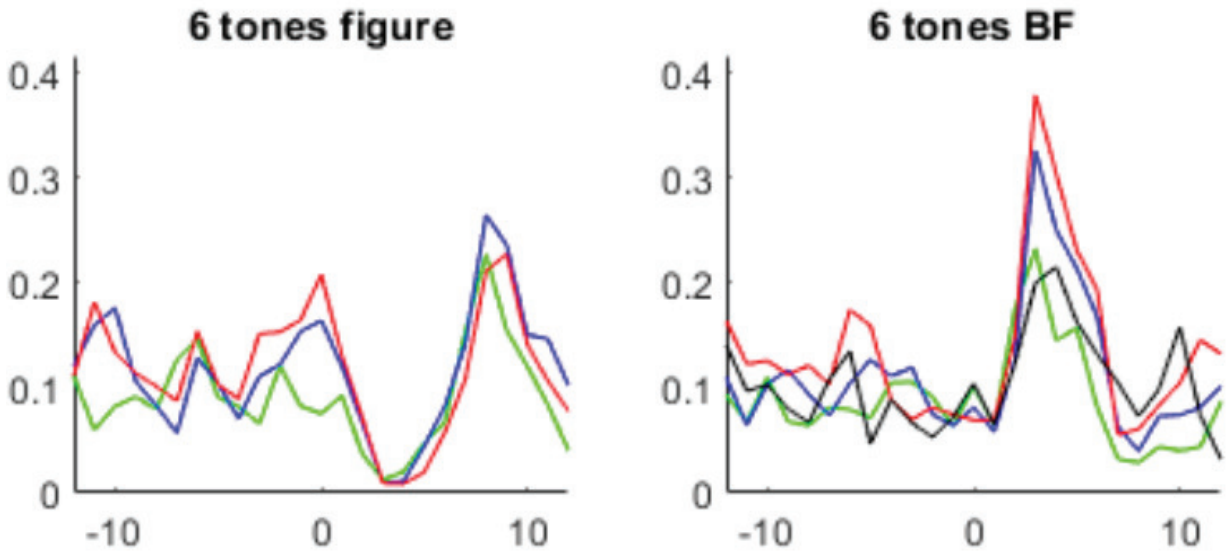


Figure 4.9: Figure and BF -triggered PSTH during the early (green), middle (blue), and late red) portions of the figure for one cell. The BF was chosen based on responses during the pre - tone cloud. For this cell, the responses to the figure and the BF are opposite in direction so we would predict a net negative post - pre response.

4.4 Discussion

Here, we have demonstrated the use of a stochastic figure-ground stimulus to probe mechanisms of auditory streaming during passive listening. Temporally coherent tones elicited A1 response modulations in untrained ferrets that persisted after the tones became incoherent again. The modulation strength was systemically related to the coherence level (number of temporally coherent tones) and primarily excitatory. This modulation response persistence and the slow time course of modulation buildup suggest that streaming-like processes are occurring even in the passive listening animal.

Immediately apparent questions raised by these results are (1) what role does attention play in response modulations to TC stimuli? and (2) how sensitive or tolerant is the cortex to various

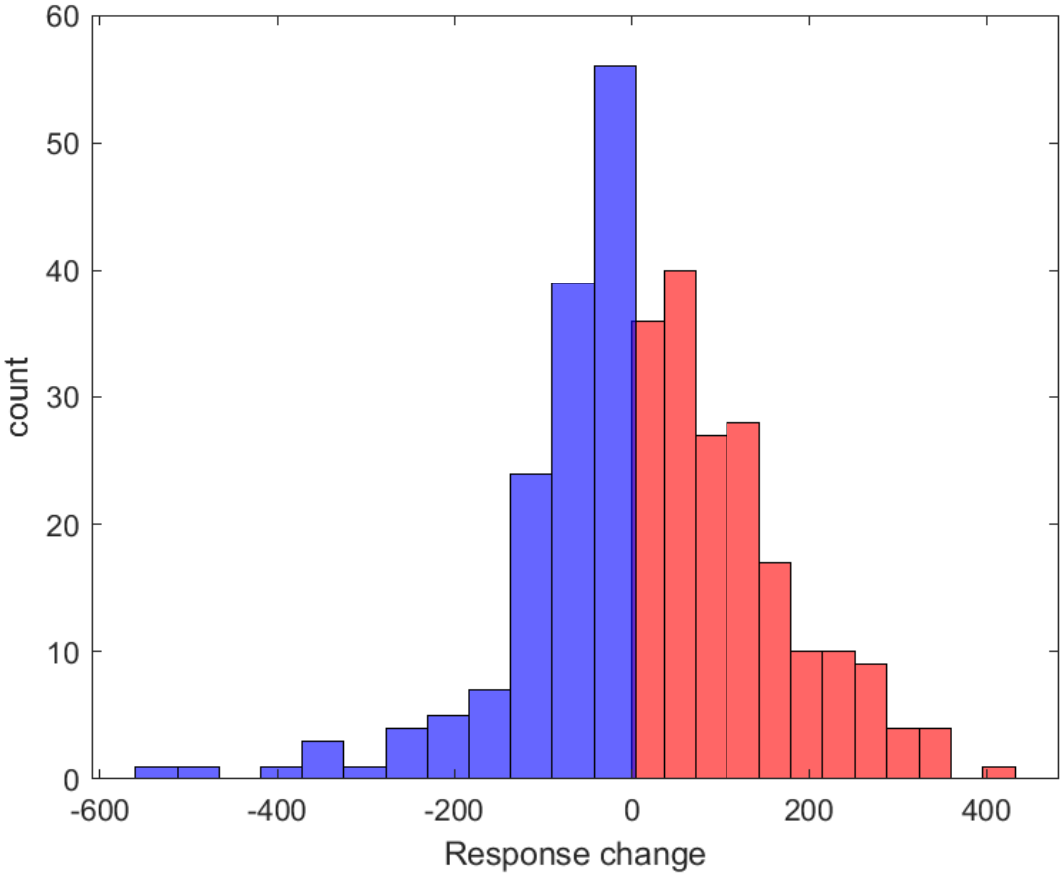


Figure 4.10: Predicting the modulatory response from initial tuning of 3 animals. Blue bars indicate the population of neurons with negative correlation between BF- and figure-driven responses. Red bars are the population with positive correlation. These populations are completely segregated

types of temporal coherence? The former question may be addressed by carefully implementing a figure-detection task or global attention with SFG stimuli in ferrets. The latter suggests a number of interesting stimulus modifications, including jittering the figure onsets, modifying the spectrotemporal density of the tone cloud, and eventually investigating the effects of including harmonically related tones in the figure.

This stimulus has rich potential for future studies of auditory processing because of its versatility. It can be generated with various spectral spread, spectrotemporal densities, and figure tone number and spacing. Johns et al. [136] used an almost identical stimulus with human subjects and showed a correlation in performance on a figure detection task and speech-in-noise comprehension. Mechanisms of streaming formation and segregation processes are open questions in addressing age-related hearing loss and central auditory processing disorder. The SFG has the advantage of avoiding non-acoustic cues (i.e. linguistic context) that can assist humans in performing speech streaming tasks, while also facilitating cross-species comparisons in brain processing strategies.

Chapter 5: Selective Attention Streaming

5.1 Introduction

As described in previous sections, TC has been implicated in streaming processes by a number of recent behavioral and human brain studies, however few studies have explored this topic at a neuronal level. The body of literature describing the perception of alternating pure tones can be summarized as follows; tones are more likely to fuse and be perceived as a single stream when they are presented at faster repetition rates and when they are separated by smaller Δf. These two factors also increase the likelihood that two synchronously presented tones are fused into one stream, and are even tolerant to small fluctuations in tone frequencies [142]. The temporal coherence hypothesis suggests that the underlying mechanism behind tone fusion involves coincidence computations behind performed between neurons across the frequency spectrum. Neurons with correlated responses co-enhance and form adaptive, excitatory relationships, while neurons with uncorrelated responses mutually suppress. The role of attention in this process is to provide a selection criterion when one or more streams are present; thus any and all inputs coincident with an attended sound inherit attentional enhancement via mutual excitation.

Lu et al. [53] explored the cortical activity in response to alternating and synchronous tone presentation, using ferrets trained on a global attention task (See Sec. 1.3 for a comparison of global and selective attention). They found that synchronous presentation resulted in

STRF enhancement during behavior relative to the passive state, supporting the TC hypothesis that synchronous tones are perceptually, or neurally, bound together independent of attention. However, the global attention task did not require any stream selection by the animal and thus we hypothesize that the modulatory effect of TC was smaller than it would be in, for example, a cocktail party scenario. In this study, we attempted to implement a selective attention task in which animals attend to one periodic stream *while ignoring* a tone presented in a non-overlapping fashion. Animals are trained to attend to a noise stream, while synchronous and asynchronous pure tones are presented as potential distractors. We hypothesized that selective attention to one stream will bind coherent tones that are synchronous with the noise. Coherent tones will be enhanced because they will inherit attentive-dependent plasticity via this binding, while alternating, incoherent tones will be suppressed.

5.2 Methods

Female ferrets are trained on an appetitive, selective attention streaming task to detect an auditory intensity cue. Animals are placed on water restriction and during task performance, must withhold licking from a water spout until the target sound is detected. The sound stimulus consists of alternating pure tones of 80ms duration with a cosine ramp of 5ms and 80ms gap of silence between tones. A narrowband noise(width 1/2 octave) burst plays synchronously(coherent) with one of the pure tones. All reference tones and noise are at a constant intensity until the target occurs. After 4-30 alternating burst epochs, the target intensity cue occurs and the animal may lick to obtain a water reward. The target is a single noise burst played 15-20 dB more intense than during the reference portion.

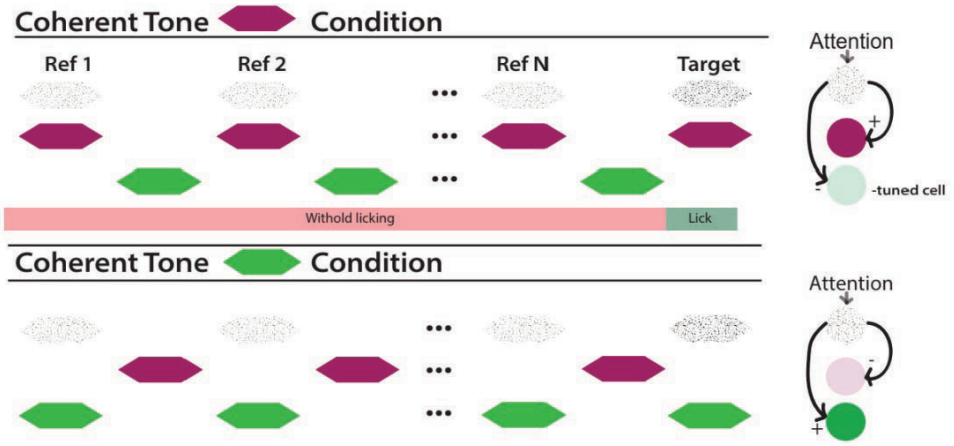


Figure 5.1: *left* Schematic of the intensity difference detection selective attention task. Reference number N = 10-30 *right* Hypothesized cell interaction based on TC theory. Attention to the noise (located $>$ 1 octave from pure tones) results in relative suppression of the asynchronous tone and enhancement of the synchronous tone.

Two pure tone frequencies A and B were chosen based on the tuning properties of the neuronal units measured at a given recording depth (see Sec 3.4). Effort was made to choose frequencies near one or more units' BFs that would produce distinct responses in all units. Animals performed the task in 2 conditions; with tone A synchronous with the noise stream (*A-SYN*) and with tone B synchronous with the noise stream (*B-SYN*). The number of references bursts prior to the target numbered between 4 and 30, with the exact count varied by task and pseudo-randomly trial to trial. Initially, animals performed all trials of each A-SYN and B-SYN conditions in two consecutive blocks, but later experiments randomly interspersed the two conditions. Animals must respond within 1 second after the target onset in order to obtain a reward. Because the response window is longer than the target sound, 5 additional references are played after the target. Post-target reference sounds ensure that the animal is rewarded for detecting the target sound, not simply the silence signaling the end of the trial.

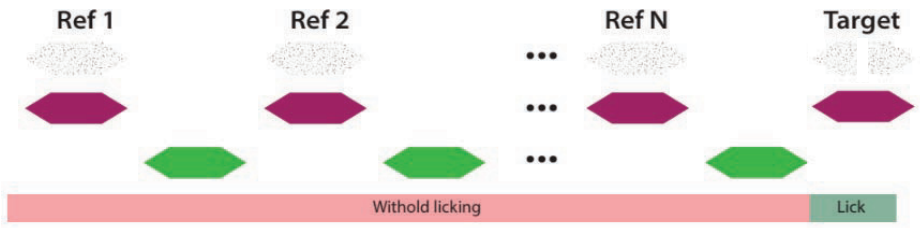


Figure 5.2: Schematic of the gap detection selective attention task. Reference number N = 4-10

Within the pre-trial reference set, tones were divided into 5-burst bins for lick count calculation. Animals were allowed to lick during up to 50% of bins without triggering a terminated trial. Binning in this way prevents the animal from using a sampling strategy, that is licking once every few references in order to trigger a reward without actually listening and detecting the target. Animals were trained until they reached at least 50% discrimination rate and 50% difference in lick rate at the short trial target time. That is, if the reference counts for a trial block were 10 and 20 references, during the 20-reference trials the animal licked at the 10-reference point less than 50% of the time. This metric is to ensure that the animal is not using a timing cue and simply waiting a fixed period of seconds before licking.

Pure tone frequencies were chosen for each experiment based on the response properties of the neurons being recorded. Ideally, tones are chosen so that one but not both tones falls near the best frequency of each neuron or recording site. The narrowband noise is placed with center frequency at least one octave above the higher pure tone.

In many experiments, noise-only trials are also played in the passive and active conditions. This paradigm represents what we hope to be a streaming task because the animal is trained to attend to the noise, while both pure tones are distractors. However, we hypothesized that the coherent tone would be perceptually "bound" with the attended noise and therefore neurons

responding to this tone would have enhanced responses. Analogously, neurons tuned to the asynchronous or non-coherent tone were hypothesized to be inhibited.

In the intensity cue task, all 3 animals were trained until reaching threshold performance, defined as 50% discrimination rate and 50% difference in lick rate at the short trial target time (See 3.1). False-alarm trials resulted in a terminated trial and time-out session. For animals 1 and 2, after this point the false-alarm trials were not terminated, and the animal was allowed to receive a reward if they successfully licked during the response window. For animal 3, false-alarm trials continued to be terminated early throughout recording. The former two animals achieved average hit rates of 93.30% $\pm$ 9.9% and 90.01% $\pm$ 11.87% with false alarm rates of 52% and 41%, respectively. The third animal achieved an average hit rate of 72.6% $\pm$ 18.86%.

In the gap detection task, both animals (4 and 5) were trained until reaching threshold performance before data collection, with false-alarm trials resulting in a terminated trial and time-out session. Animals 4 and 5 achieved hit rates of 70.91% $\pm$ 29.22% and 51.57% $\pm$ 17.97% , respectively. Poor performance by animal 5 may be explained by the transition from training to data collection resulting in a much longer waiting period in the holder.

5.3 Results

5.3.1 Intensity Cue

Single-unit neuronal responses were sorted and binned into 10 ms windows and averaged over each 320 ms epoch of the stimulus (2, 80 ms tones separated by 80ms gaps) to obtain peri-stimulus time histograms (PSTH), as shown in figure 5.3. Most of the units came from animals 1 and 3 (98 and 99 units, respectively, and the remaining 42 from the 2nd animal). Responses

were analyzed beginning at the middle third of reference epochs because we assume that it may require a few repeats for stream formation to occur. Synchronous (SYNC) and alternating (ALT) conditions for each neuron were assigned by comparing average spike rate during passive presentation of the two alternating tone stimuli. Tone A response values were calculated by summing tone A+noise (A-ASYN, first half) with tone A alone (B-SYN, second half) responses. The same procedure was followed for tone B response values. Stimuli were assigned as SYNC or ALT relative to each cell based on which of these responses values was greater. For example, if a neuron spikes more during (A+noise+A alone) than (B+noise+B alone), we consider A-SYN to be the SYNC stimulus for that cell and B-SYN to be the ALT stimulus. Only neurons with a minimum 10% difference in spike rate were considered. This selection also ensured that neurons driven only by the noise stream and not tone A or B were removed.

Many neurons showed changes in responses during the 4 experimental conditions (SYNC passive, ALT passive, SYNC active, ALT active), as exemplified in figures 5.4a and 5.4b. As expected, firing rates were globally suppressed during attention, with enhancement of neurons tuned to the synchronous frequency.

Histograms comparing spike rates in 4 experimental conditions for all 3 animals are shown in figure 5.5 with overlaid gamma distribution fit lines. A Wilcoxon signed rank test was performed to determine significant differences between mean firing rates during the 4 conditions, using the Bonferroni correction for multiple comparisons. For animals 1 and 2, firing rates in the SYNC active condition was significantly higher than the other three distributions ($p < 0.05$). These significance relationships held even when false alarm trials from animals 1 and 2 were included in the analysis. No other pairwise comparison was statistically significant. This is particularly notable because we initially selected units with at least a 10% difference in response to

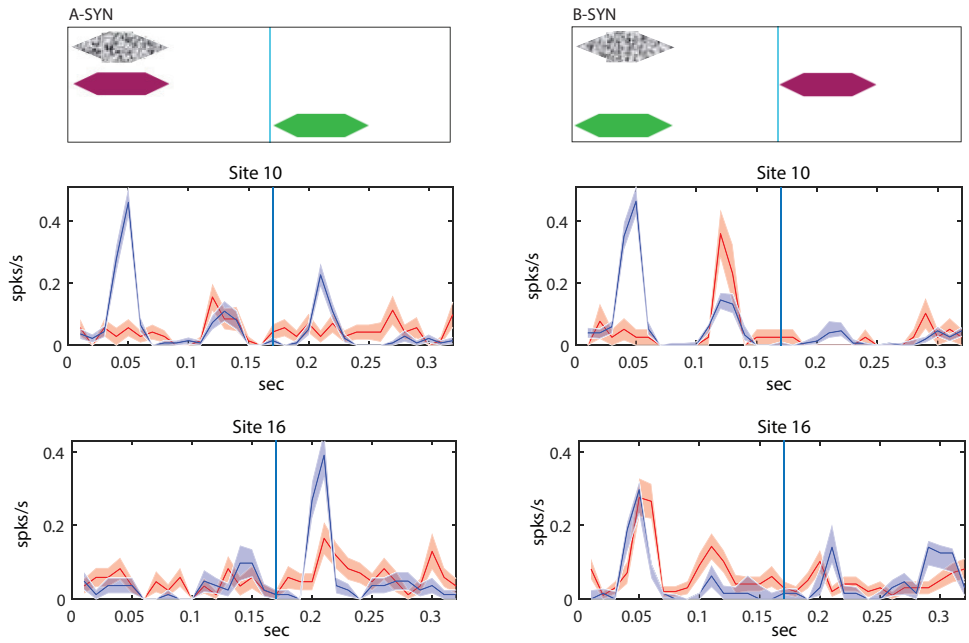
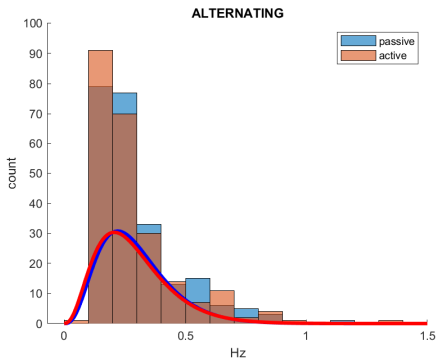


Figure 5.3: Two neurons exemplify TC and attention - modulated responses. *top row* diagram of A-SYN and B-SYN epochs for PSTH window alignment. Responses are folded over the length of one tone cycle, with noise onset at t = 0 and asynchronous tone onset at t = 0.16 s. *middle and bottom rows* Example neuron responses to each experimental condition with passive responses in blue, active responses in red. Left and right plots are A-SYN and B-SYN, respectively.

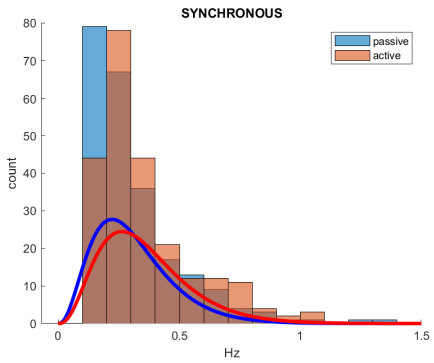
(a) Synchronous vs. alternating

(b) Active vs. passive

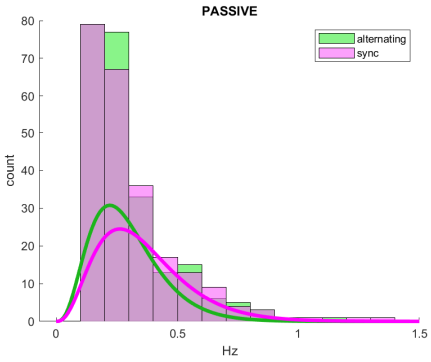
Figure 5.4: Comparison of firing rate in 4 experimental conditions for intensity cue task. *left* Alternating responses are essentially symmetric for both active and passive presentation, while synchronous responses skew above the diagonal. *right* Active responses are heavily skewed above the diagonal, indicating an increase in firing rate for synchronous over alternating behavior. Passive responses are symmetric.



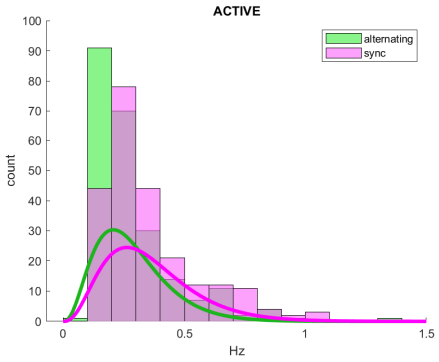
(a) Alternating, active vs. passive



(b) Synchronous, active vs. passive



(c) Passive, alternating vs. synchronous



(d) Active, alternating vs. synchronous

Figure 5.5: Histograms of intensity detection spiking conditions with fit curves. There is a significant rightward shift in plots (b) and (d). Plot (c) appears to have a shift, but it is not statistically significant.

tones A and B. However, the passive responses to SYNC and ALT stimuli were not statistically significant. Therefore, the rightward shift of the distributions can attributed entirely to attentional engagement.

5.3.2 Gap detection

Data from the gap detection experiment was analyzed in the same manner as above, but with PSTH windows adjusted to reflect the adjusted stimulus parameters. Specifically, one reference repeat for the intensity cue sound had duration 320 ms, while one repeat for gap detection had duration 620 ms. Animals 4 and 5 contributed 39 and 49 neuronal units, respectively. Few

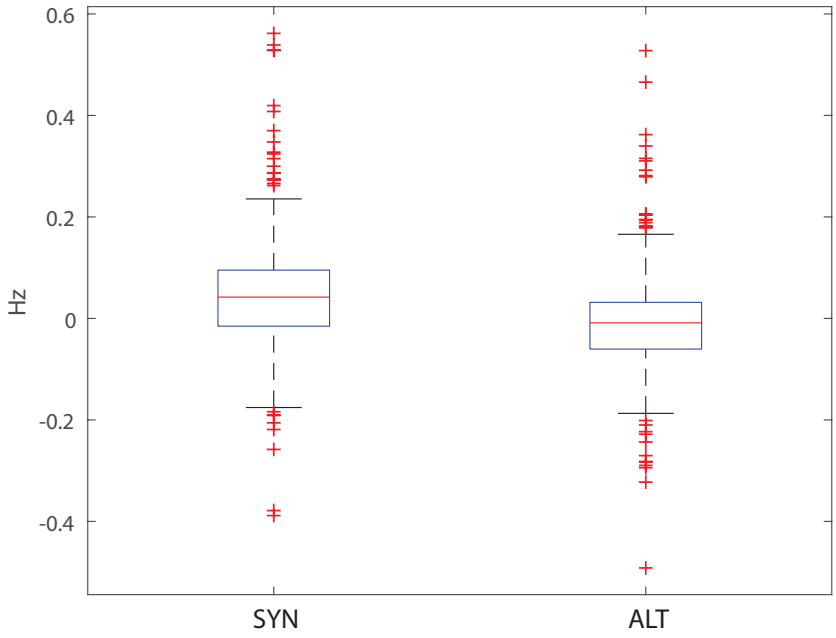
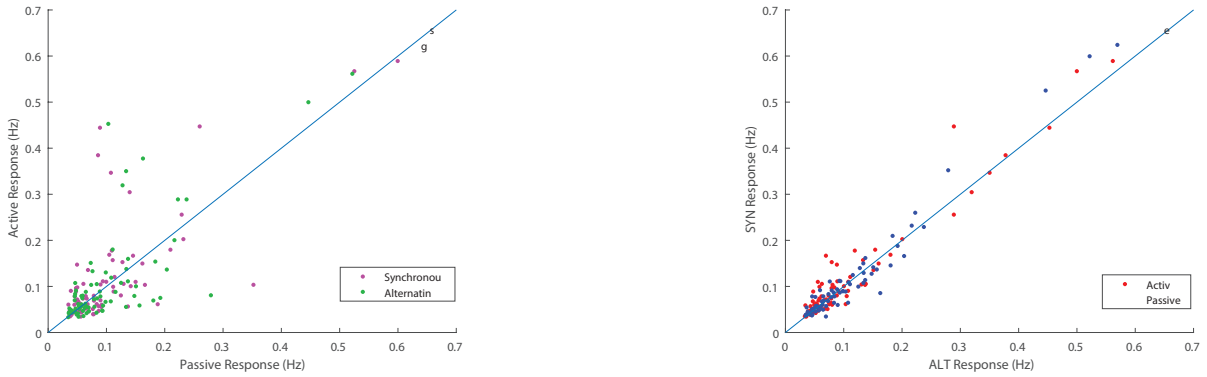


Figure 5.6: *left* The per-unit difference between synchronous active and passive has a positive mean and positive-leaning distribution. *right* The difference between active and passive in alternating condition is centered at 0. These two distributions are significantly different.



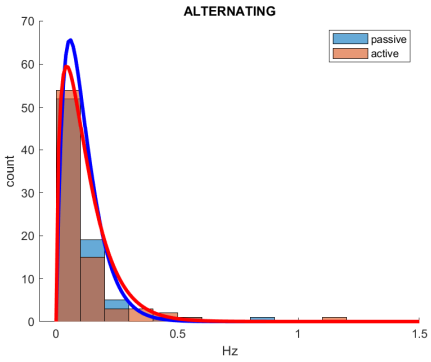
(a) Synchronous vs. alternating

(b) Active vs. passive

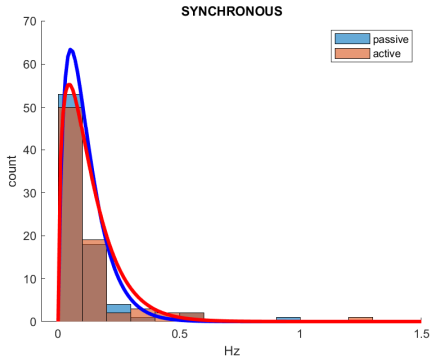
Figure 5.7: Comparison of firing rate in 4 experimental conditions for gap detection task in both animals. *left* Alternating and synchronous responses are essentially symmetric for both active and passive presentation. *right* Active responses are slightly skewed above the diagonal, due to the contribution of animal 4, indicating an increase in firing rate for synchronous over alternating behavior. Passive responses are symmetric.

neurons showed changes in responses to synchronous vs. asynchronous conditions in the behaving case (See figs. 5.7a and 5.7b). Here, the timing of the PSTH is different because we used longer tones and smaller silences between tones in order to accommodate a gap in the target. Accordingly, we reduced the number of reference bursts to 4 to 10 per trial in order to keep the overall trial length comparable to the intensity cue task. On a population level, neither animal showed a significant difference between SYNC active and ALT active ($p < 0.05$).

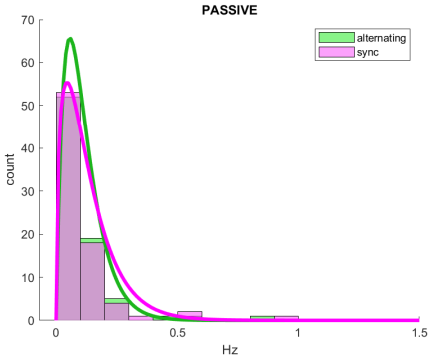
We hypothesize that the poor behavioral performance paired with small data set prevented a robust effect from being evident in the gap detection animals. However upon reflection, gap detection was not the best choice for a streaming tasks. Because the gap interrupts the regular periodicity of the tone streams, animals may be able to rely on passive mismatch negativity mechanisms in order to detect the deviant (see Sec. 1.2.2) for a summary of temporal regularity in streaming). It unclear if that was the case for these animals, but it is something to keep in mind for interpreting results and a strong caveat to continuing with this stimulus.



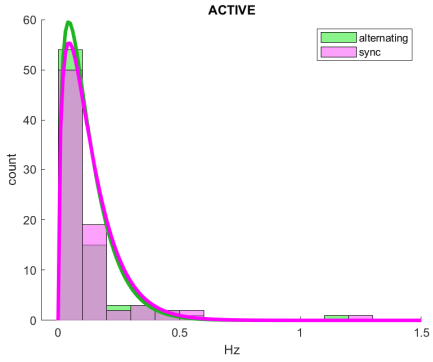
(a) Alternating, active vs. passive



(b) Synchronous, active vs. passive



(c) Passive, alternating vs. synchronous



(d) Active, alternating vs. synchronous

Figure 5.8: Histograms of gap detection spiking conditions with fit curves. Only (d) demonstrates a significant change.

5.4 Discussion

The intensity cue selective attention streaming tasks demonstrated a difference in neuronal responses for units tuned to frequencies which were synchronous vs. asynchronous with the attended stream. In particular, we showed that the combination of active, attended listening and synchronous presentation with an attended stream significantly modulated neuronal firing rates in a way that is not produced by either of those conditions individually. Our results support the hypothesis that selective attention causes enhancement of responses to other sources temporally coherence with the attended stream.

However, framing these tasks as selective attention may be a generous interpretation. In fact, it is theoretically possible to perform both tasks as deviance detection rather than streaming tasks. It is not possible to examine the "mind of the animal" and determine what strategy it is employing to solve a task, but global attention or even inattention may explain why including false-alarm trials did not extinguish the significance relationships between the 4 experimental conditions for animals 1 and 2. It may also contribute to explaining why animal 4 had good performance but showed no significant difference in spiking data, while animal 5 had poor performance but some significant differences. If the animals did employ global rather than selective attention, we have at least reproduced the results of Lu et al. [53], that tone synchrony can induce a small but significant enhancement in neuronal firing.

Regardless of the attentional strategy employed by the animals, we have demonstrated that when both synchronous and alternating tones are present, the synchronous tones will bind and mutually enhance while alternating tones are unmodulated. This result builds on those of Lu et al. [53], who showed that when *only* synchronous **or** alternating tones are present, synchronous

tones will be relatively enhanced. Our result is highly significant because it shows differential treatment of SYNC and ALT within-trial and during concurrent presentation.

In order to ensure selective attention, the task must include two adversarial streams and a target which cannot be detected via global listening or deviance detection. One option would be to use a two-speaker sequence detection task like that used by Joshi [12]. The animal listens for a target word uttered by a specific speaker, and the background speaker occasionally uses the target word but must be ignored. Once the animal can perform this task (no easy feat on its own), new tones could be introduced which are inharmonic but temporally coherent with one of the voices. This would allow us to probe whether TC of inharmonic tones will drive responses similar to harmonics of the voice stream.

A key factor in these experiments is that the noise stream was placed spectrally distant from the pure tones in order to prevent contamination of responses. Attention is commonly described as a spotlight which may "shed light" on an area of focus, but may shed leaky enhancement effects to nearby areas, i.e. frequencies. In this experiment, the spacing of the tones ensured that they were not driven by the noise during passive presentation and suggests that the observed effects were due to TC, and not proximity to the attended stream.

Chapter 6: Conclusions

The neuronal mechanisms underlying auditory perception are largely unknown, with a primary mystery being how the brain fuses disparate frequency components into unified sound percepts. In this dissertation work, I present three studies, two investigating the role of temporal coherence across frequencies to induce strean formation, and one postulating a pitch-neuron model in the early auditory system. These results present evidence that the auditory system is sensitive to TC in both attended and unattended states, but the role of attention and salience may be important in mediating streaming processes.

In chapter 2, we postulated a biophysical model for how pitch could be computed in the early auditory system. The existence of such a neuron has been posited before, but never with such detail as to how the computation could be performed. Recent work has shown that pitch encoding exists in the auditory cortex, which we assume is inherited from an earlier stage due to the necessity of phase-locking for resolved pitch. After poring over the literature on brainstem connectivity and the limited available information on physiology in that region, we strongly suspect that these cells may exist in the VNLL. The validation of this model would not only solve a landmark question in auditory perceptual mechanisms, but strongly emphasize the importance of dendritic morphology and resonances in understanding how neurons integrate inputs, especially periodic ones. While spike-timing dependent plasticity has been a popular mechanism for some

time now, the addition of dendritic filtering and lag could be a crucial part of the puzzle of how coincidence detection or correlation-like operations are performed in the brain. These represent a class of operations that we have long suspected are taking place in the brain, but have never found the elaborate network structure expected to perform them.

Chapters 4 and 5 explore the neural correlates of temporal coherence from two different angles. The stochastic figure ground experiment used passive animals and relied on an assumption that a sufficiently salient figure will create a perceptual pop-out and corresponding neural response, even in an inattentive listener. This is a reasonable assumption, given that passive listening ferret brain clean noisy speech [29]. Indeed, we found effects of temporal coherence in A1 in 2 of 3 animals. It's likely that attention would increase this effect, particularly in higher-order areas known to exhibit stronger attention-dependent plasticity. Given that different parameters for a tone cloud can produce different response field estimate, our results may be strengthened or diminished by a different choice of parameters. It's difficult to choose an optimal set for ferrets without first performing a behavioral study, but we have at least shown that an effect exists, with potential for further investigation.

Chapter 5 approached TC from the perspective of stimulus-stimulus pairing where one stimulus (the noise) is task relevant and the other (the tones) is uninstructed and irrelevant [143]. We hypothesized that attention can unintentionally leak from an attended tone to an unattended one via temporal coherence. Here, we tried to trick the animal into streaming. Our results met with mixed success, with significant changes in firing rate in 3 of 5 animals, and the other two animals possibly explainable by poor behavior or choice of task structure. This project built on the work of Lu [53], who suggested that selective attention would produce stronger effects in A1 than global attention with synchronous tones. In retrospect, our experimental design did not

require selective attention and could be performed using a global attention strategy. So, Lu's hypothesis will take another experiment to test.

These results suggest both top-down (task-driven) and bottom-up (saliency-driven) processes have a role to play in streaming. We suggest that pitch, which is a strong cue for harmonic cancellation of background sounds [de Cheveigne 2020], is computed in the brainstem or midbrain at the latest, well upstream of AC. Furthermore, we observed a consistent enhancement of neural responses only during the attending, synchronous condition of the selective attention task, while the figure-ground study yielded strong effects with no directed attention. While AC is strongly implicated in both of these mechanisms [Elhilali 2009], it is difficult to say whether our observed effects in A1 are driven by modulations in brain regions that are up- or downstream of the AC.

Throughout this dissertation we have argued that the effects of TC in stream formation are demonstrated through modulation of auditory cortical responses to sounds which are precisely synchronized vs. those that are spectrotemporally non-overlapping and/or randomly occurring. We argue that the effects seen are evidence of TC because of the time course of the modulation; several seconds of listening are required to observe the largest differences in cortical responses. This is much slower from what we would expect if the animal was listening specifically to detect a specific frequency or set of frequencies [9, 10]. Our results differ notably from previous streaming studies in that the primary observed effect was enhancement or increase in neural responses to TC tones. Previous studies have primarily demonstrated suppression of competing sounds [9, 12, 39, 144], however some have found primarily relative enhancement [42, 50, 59]. One possible cause of this difference is that our observations focused on A1, while non-primary auditory cortex has been implicated in suppression during complex listening in humans [145]. At-

tentional modulation of neural responses in ferrets is stronger in A2 and tertiary areas [144, 146], but has not been studied specifically with respect to streaming processes. Furthermore, recent work has suggested that human A1 may be analogous to the ferret A2 [12, 147], therefore the animal A1 may be analogous to a subcortical region upstream of the human A1.

6.1 Future Work

The implications of these studies for future lines of research are numerous and branching. First, the pitch model as described in Chapter 2 suggests candidate locations in the brainstem for such a neuron to exist. Validation of this model using *in vivo*, intracellular recording methods would be an invaluable starting point for further investigation into the neural substrates of pitch and harmonicity processing. While we strongly believe that the initial computation of perceptual pitch must occur prior to the midbrain, harmonic-sensitive regions have been found in secondary AC, suggesting that pitch encoding is preserved throughout the auditory pathway. Biological validation and localization of our model would no doubt inform future studies of this harmonicity sensitive encoding.

These insights into harmonicity and spectral cues for streaming would further open up investigation into the interplay between harmonicity and temporal coherence, and the related phenomena described by Micheyl [142]. If pitch is indeed computer in the brainstem and encoded in the secondary cortex, then presumably it must be encoded all along the hierarchy and we have missed it all this time. Another hypothesis is that the pitch computation operation is re-learned in the cortex through long-term exposure to harmonic sounds. In this case, there may be a disjoint between perceptual pitch and cortical pitch. The SFG stimulus as we have formulated it could be

used for such investigations, by simply choosing a more spectrally dense frequency set to include multiple harmonics per octave.

Our results using the temporally coherent tone cloud stimulus also raise many questions primarily in two directions (1) is attention necessary for streaming processes, and (2) what are the limits of temporal coherence to induce stream formation. The former question is difficult to address objectively, as we cannot be sure that a passive animal isn't subconsciously listening, however the answer seem to be "not all the time." Combining the results of chapters 4 and 5 suggests that attention plays a role, but it may depend on the complexity of the stimulus or the specific task underlying attentive listening. So, if attention is not necessary, how salient must the stream be in order for it to form? Or how temporally coherent should it be? The latter question of course is identical to (2). The next logical step for these studies is to implement a small amount of jitter in the figure onsets. It is interesting to consider whether there is any perceptual difference between systematic jitter, fixed jitter, or random jitter; are adaptive coherence relationships between cells flexible to lag? Bidirectional lag? We know from alternating and synchronous tone studies that perceptual fusion of 2 tones is tolerant to some degree of lag, but will a more naturalistic stimulus produce the same results? Whatever the case may be, this line of questioning is sure to be fruitful for generations of scientists, with the SFG as a convenient stimulus in the toolbox.

Bibliography

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953. [Online]. Available: /record/1954-04260-001

[2] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 8 2005. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1945807

[3] A. J. Oxenham, "Pitch perception," *Journal of Neuroscience*, vol. 32, pp. 13 335–13 338, 9 2012. [Online]. Available: https://www.jneurosci.org/content/32/39/13335https://www.jneurosci.org/content/32/39/13335.abstract

[4] A. J. M. Houtsma and J. Smurzynski., "Pitch identification and discrimination for complex tones with many harmonics," *Journal of the Acoustical Society of America*, vol. 87, pp. 304–310, 1 1990. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.399297

[5] B. K. Lau, A. H. Mehta, and A. J. Oxenham, "Superoptimal perceptual integration suggests a place-based representation of pitch at high frequencies," *Journal of Neuroscience*, vol. 37, pp. 9013–9021, 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28821642/

[6] A. J. Oxenham, C. Micheyl, M. V. Keebler, A. Loper, and S. Santurette, "Pitch perception beyond the traditional existence region of pitch," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 7629–7634, 5 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21502495http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3088642http://www.pnas.org/cgi/doi/10.1073/pnas.1015291108

[7] F. A. Rodríguez, C. Chen, H. L. Read, and M. A. Escabí, "Neural modulation tuning characteristics scale to efficiently encode natural sound statistics." *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, pp. 15 969–80, 11 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21106835http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3351116

[8] M. Elhilali, J. B. Fritz, T. S. Chi, and S. A. Shamma, "Auditory cortical receptive fields: Stable entities with plastic abilities," *Journal of Neuroscience*, vol. 27, pp. 10 372–10 382, 9 2007. [Online]. Available: https://www.jneurosci.org/content/27/39/10372https://www.jneurosci.org/content/27/39/10372.abstract

[9] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature Neuroscience*, vol. 6, pp. 1216–1223, 11 2003. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/14583754/

[10] S. Atiani, M. Elhilali, S. V. David, J. B. Fritz, and S. A. Shamma, "Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields," *Neuron*, vol. 61, pp. 467–480, 2 2009.

[11] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Adaptive changes in cortical receptive fields induced by attention to complex sounds," *Journal of Neurophysiology*, vol. 98, pp. 2337–2346, 10 2007.

[12] N. H. Joshi, "Speech segregation and representation in the ferret auditory and frontal cortices," Ph.D. dissertation, University of Maryland, College Park, MD, Oct 2022.

[13] N. Ulanovsky, L. Las, D. Farkas, and I. Nelken, "Multiple time scales of adaptation in auditory cortex neurons," *Journal of Neuroscience*, vol. 24, pp. 10 440–10 453, 11 2004. [Online]. Available: https://www.jneurosci.org/content/24/46/10440https://www.jneurosci.org/content/24/46/10440.abstract

[14] N. Itatani and G. M. Klump, "Animal models for auditory streaming," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, p. 20160112, 2 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28044022http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5206279http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2016.0112

[15] H. Cai and M. L. Dent, "Attention capture in birds performing an auditory streaming task," *PLOS ONE*, vol. 15, p. e0235420, 6 2020. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0235420

[16] K. L. Christison-Lagay and Y. E. Cohen, "Behavioral correlates of auditory streaming in rhesus macaques," *Hearing research*, vol. 309, pp. 17–25, 3 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24239869/

[17] K. M. M. Walker, J. W. H. Schnupp, S. M. B. Hart-Schnupp, A. J. King, and J. K. Bizley, "Pitch discrimination by ferrets for simple and complex sounds," *The Journal of the Acoustical Society of America*, vol. 126, p. 1321, 9 2009. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.3179676

[18] L. Ma, C. Micheyl, P. Yin, A. J. Oxenham, and S. A. Shamma, "Behavioral measures of auditory streaming in ferrets (mustela putorius)," *Journal of comparative psychology (Washington, D.C. : 1983)*, vol. 124, p. 317, 8 2010. [Online]. Available: /pmc/articles/PMC4312586//pmc/articles/PMC4312586/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4312586/

[19] D. Saderi, B. N. Buran, and S. V. David, "Streaming of repeated noise in primary and secondary fields of auditory cortex," *Journal of Neuroscience*, vol. 40, pp. 3783–3798, 5 2020. [Online]. Available: https://www.jneurosci.org/content/40/19/3783https://www.jneurosci.org/content/40/19/3783.abstract

[20] A. S. Bregman, *Auditory scene analysis : the perceptual organization of sound*. MIT Press, 1990.

[21] T. D. Griffiths and J. D. Warren, "What is an auditory object?" *Nature Reviews Neuroscience 2004 5:11*, vol. 5, pp. 887–892, 11 2004. [Online]. Available: https://www.nature.com/articles/nrn1538

[22] A. de Cheveigné, "Harmonic cancellation-a fundamental of auditory scene analysis," *Trends in hearing*, vol. 25, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34698574/

[23] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as bayesian inference," *Annual review of psychology*, vol. 55, pp. 271–304, 2004. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/14744217/

[24] N. Saint-Arnaud and K. Popat, *Analysis and synthesis of sound textures*. CRC Press, 1998, pp. 293–308. [Online]. Available: http://130.203.136.95/viewdoc/summary?doi=10.1.1.111.586

[25] W. M. Hartmann, S. McAdams, and B. K. Smith, "Hearing a mistuned harmonic in an otherwise periodic complex tone," *The Journal of the Acoustical Society of America*, vol. 88, pp. 1712–1724, 1990. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/2262628/

[26] B. C. Moore, B. R. Glasberg, and R. W. Peters, "Thresholds for hearing mistuned partials as separate tones in harmonic complexes," *The Journal of the Acoustical Society of America*, vol. 80, pp. 479–483, 1986. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/3745680/

[27] S. Popham, D. Boebinger, D. P. Ellis, H. Kawahara, and J. H. McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature Communications 2018 9:1*, vol. 9, pp. 1–13, 5 2018. [Online]. Available: https://www.nature.com/articles/s41467-018-04551-8

[28] M. J. McPherson, R. C. Grace, and J. H. McDermott, "Harmonicity aids hearing in noise," *Attention, perception I& psychophysics*, vol. 84, pp. 1016–1042, 4 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35102502/

[29] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Mechanisms of noise robust representation of speech in primary auditory cortex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 6792–6797, 5 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24753585/

[30] A. de Cheveigné, "Ten experiments on vowel segregation," *ATR Human Information Processing Research Labs technical report TR-H-217*, 1997. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03090891https://hal.archives-ouvertes.fr/hal-03090891/document

[31] R. L. Freyman, A. M. Griffin, and A. J. Oxenham, "Intelligibility of whispered speech in stationary and modulated noise maskers," *The Journal of the Acoustical Society of America*, vol. 132, p. 2514, 10 2012. [Online]. Available: /pmc/articles/PMC3477190//pmc/articles/PMC3477190/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3477190/

[32] D. Bendor and X. Wang, "The neuronal representation of pitch in primate auditory cortex," *Nature*, vol. 436, pp. 1161–1165, 8 2005. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16121182/

[33] E. J. Allen, J. Mesik, K. N. Kay, and A. J. Oxenham, "Distinct representations of tonotopy and pitch in human auditory cortex," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 42, pp. 416–434, 1 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34799415/

[34] N. Suga, W. E. O'Neill, and T. Manabe, "Harmonic-sensitive neurons in the auditory cortex of the mustache bat," *Science*, vol. 203, pp. 270–274, 1979. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/760193/

[35] C. Brodbeck and J. Z. Simon, "Cortical tracking of voice pitch in the presence of multiple speakers depends on selective attention," *Frontiers in Neuroscience*, vol. 16, p. 1224, 8 2022.

[36] E. Sohoglu and M. Chait, "Detecting and representing predictable structure during auditory scene analysis," *eLife*, vol. 5, 9 2016.

[37] L. Aman, S. Picken, L. V. Andreou, and M. Chait, "Sensitivity to temporal structure facilitates perceptual analysis of complex auditory scenes," *Hearing Research*, vol. 400, p. 108111, 2 2021.

[38] I. C. Griffin, C. Miniussi, and A. C. Nobre, "Orienting attention in time," *Frontiers in bioscience : a journal and virtual library*, vol. 6, 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11282565/

[39] M. Chait, A. de Cheveigné, D. Poeppel, and J. Z. Simon, "Neural dynamics of attending and ignoring in human auditory cortex," *Neuropsychologia*, vol. 48, pp. 3262–3271, 9 2010.

[40] R. Southwell, A. Baumann, C. Gal, N. Barascud, K. Friston, and M. Chait, "Is predictability salient? a study of attentional capture by auditory patterns," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, 2 2017. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0105

[41] L. V. Andreou, M. Kashino, and M. Chait, "The role of temporal regularity in auditory segregation," *Hearing Research*, vol. 280, pp. 228–235, 10 2011.

[42] M. Elhilali, L. Ma, C. Micheyl, A. J. Oxenham, and S. A. Shamma, "Temporal coherence in the perceptual organization and cortical representation of auditory scenes," *Neuron*,

vol. 61, pp. 317–329, 1 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19186172http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2673083http://linkinghub.elsevier.com/retrieve/pii/S0896627308010532

[43] L. Krishnan, M. Elhilali, and S. Shamma, "Segregating complex sound sources through temporal coherence," *PLoS Computational Biology*, vol. 10, p. 1003985, 12 2014. [Online]. Available: /pmc/articles/PMC4270434/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270434/

[44] S. K. Christiansen, M. L. Jepsen, and T. Dau, "Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in "primitive" auditory stream segregationa)," *The Journal of the Acoustical Society of America*, vol. 135, p. 323, 1 2014. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.4845675

[45] A. S. Bregman and S. Pinker, "Auditory streaming and the building of timbre." *Canadian journal of psychology*, vol. 32, pp. 19–31, 1978. [Online]. Available: /record/1979-22752-001

[46] C. Micheyl, C. Hanson, L. Demany, S. Shamma, and A. J. Oxenham, "Auditory stream segregation for alternating and synchronous tones." *Journal of experimental psychology. Human perception and performance*, vol. 39, pp. 1568–80, 12 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23544676http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3973443

[47] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends in Neurosciences*, vol. 34, pp. 114–123, 3 2011.

[48] B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Adaptation of the human auditory cortex to changing background noise," *Nature Communications 2019 10:1*, vol. 10, pp. 1–11, 6 2019. [Online]. Available: https://www.nature.com/articles/s41467-019-10611-4

[49] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral Cortex*, vol. 25, pp. 1697–1706, 7 2015. [Online]. Available: https://academic.oup.com/cercor/article/25/7/1697/457492

[50] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, pp. 11 854–11 859, 7 2012. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1205381109

[51] M. Rezaeizadeh and S. Shamma, "Binding the acoustic features of an auditory source through temporal coherence," *Cerebral cortex communications*, vol. 2, 10 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34746791/

[52] B. Skerritt-Davis and M. Elhilali, "Detecting change in stochastic sound sequences," *PLOS Computational Biology*, vol. 14, p. e1006162, 5 2018. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006162

[53] K. Lu, Y. Xu, P. Yin, A. J. Oxenham, J. B. Fritz, and S. A. Shamma, "Temporal coherence structure rapidly shapes neuronal interactions," *Nature Communications*, vol. 8, 1 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28054545/

[54] J. Lawlor, "Tracking changes in complex auditory scenes along the cortical pathway," Ph.D. dissertation, École Normale Supérieure, Paris Sciences et Lettres University, Paris, France, Oct 2018.

[55] J. Zhao, N. Al-Aidroos, and N. B. Turk-Browne, "Attention is spontaneously biased toward regularities," *Psychological Science*, vol. 24, 2013.

[56] Y. Gabay, F. K. Dick, J. D. Zevin, and L. L. Holt, "Incidental auditory category learning," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 41, 2015.

[57] F. W. Ohl and H. Scheich, "Learning-induced plasticity in animal and human auditory cortex," *Current Opinion in Neurobiology*, vol. 15, pp. 470–477, 8 2005.

[58] J. C. Dahmen, D. E. Hartley, and A. J. King, "Stimulus-timing-dependent plasticity of cortical frequency representation," *The Journal of Neuroscience*, vol. 28, p. 13629, 12 2008. [Online]. Available: /pmc/articles/PMC2663791//pmc/articles/PMC2663791/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663791/

[59] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, and C. E. Schroeder, "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"," *Neuron*, vol. 77, pp. 980–991, 3 2013.

[60] C. Micheyl and A. J. Oxenham, "Objective and subjective psychophysical measures of auditory stream integration and segregation," *Journal of the Association for Research in Otolaryngology : JARO*, vol. 11, pp. 709–724, 12 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20658165/

[61] A. D. Patel, "Music, language, and the brain," *Music, Language, and the Brain*, pp. 1–526, 12 2007. [Online]. Available: https://academic.oup.com/book/10227

[62] D. Schön, C. Magne, and M. Besson, "The music of speech: Music training facilitates pitch processing in both music and language," *Psychophysiology*, vol. 41, pp. 341–349, 5 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15102118http://doi.wiley.com/10.1111/1469-8986.00172.x

[63] A. J. Oxenham, "Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants," *Trends in Amplification*, vol. 12, p. 316, 2008. [Online]. Available: /pmc/articles/PMC2901529//pmc/articles/PMC2901529/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2901529/

[64] B. C. J. Moore, *An Introduction to the Psychology of Hearing: Sixth Edition*. Brill, 4 2012. [Online]. Available: https://brill.com/display/title/24210

[65] T. M. Shackleton and R. P. Carlyon, "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination." *The Journal of the Acoustical Society of America*, vol. 95, pp. 3529–40, 6 1994. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8046144

[66] J. G. Bernstein and A. J. Oxenham, "Pitch discrimination of diotic and dichotic tone complexes: harmonic resolvability or harmonic number?" *The Journal of the Acoustical Society of America*, vol. 113, p. 3323, 2003. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/12822804/

[67] O. Macherey, J. M. Deeks, and R. P. Carlyon, "Extending the limits of place and temporal pitch perception in cochlear implant users," *JARO: Journal of the Association for Research in Otolaryngology*, vol. 12, p. 233, 4 2011. [Online]. Available: /pmc/articles/PMC3046333//pmc/articles/PMC3046333/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3046333/

[68] B. C. Moore and H. E. Gockel, "Resolvability of components in complex tones and implications for theories of pitch perception," *Hearing Research*, vol. 276, pp. 88–97, 6 2011.

[69] O. Macherey and R. P. Carlyon, "Re-examining the upper limit of temporal pitch," *The Journal of the Acoustical Society of America*, vol. 136, pp. 3186–3199, 12 2014. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4900917

[70] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. i. pitch and pitch salience." *Journal of neurophysiology*, vol. 76, pp. 1698–716, 9 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8890286

[71] ——, "Neural correlates of the pitch of complex tones. ii. pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch." *Journal of neurophysiology*, vol. 76, pp. 1717–34, 9 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8890287

[72] R. Meddis and L. P. O'Mard, "Virtual pitch in a computational physiological model," *The Journal of the Acoustical Society of America*, vol. 120, pp. 3861–3869, 12 2006. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.2372595

[73] J. E. Rose, J. F. Brugge, D. J. Anderson, and J. E. Hind, "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey," *Journal of neurophysiology*, vol. 30, pp. 769–793, 1967. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/4962851/

[74] D. O. Kim and C. E. Molnar, "A population study of cochlear nerve fibers: comparison of spatial distributions of average-rate and phase-locking measures of responses to single tones," *https://doi.org/10.1152/jn.1979.42.1.16*, vol. 42, pp. 16–30, 1979. [Online]. Available: https://journals.physiology.org/doi/10.1152/jn.1979.42.1.16

[75] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *The Journal of the Acoustical Society of America*, vol. 102, pp. 1811–1820, 9 1997. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.420088

[76] M. Slaney and R. F. Lyon, "A perceptual pitch detector," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, pp. 357–360, 1990.

[77] J. Laudanski, Y. Zheng, and R. Brette, "A structural theory of pitch," *eNeuro*, vol. 1, pp. 33–47, 11 2014. [Online]. Available: https://www.eneuro.org/content/1/1/ENEURO.0033-14.2014https://www.eneuro.org/content/1/1/ENEURO.0033-14.2014.abstract

[78] J. L. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *The Journal of the Acoustical Society of America*, vol. 54, pp. 1496–1516, 12 1973. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1914448

[79] S. Shamma and D. Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *The Journal of the Acoustical Society of America*, vol. 107, pp. 2631–2644, 5 2000.

[80] C. J. Plack and A. J. Oxenham, "Overview: The present and future of pitch," *Pitch*, pp. 1–6, 4 2005. [Online]. Available: https://www.researchgate.net/publication/226990482_Overview_The_Present_and_Future_of_Pitch

[81] N. Grimault, C. Micheyl, R. P. Carlyon, and L. Collet, "Evidence for two pitch encoding mechanisms using a selective auditory training paradigm," *Perception I& Psychophysics 2002 64:2*, vol. 64, pp. 189–197, 2002. [Online]. Available: https://link.springer.com/article/10.3758/BF03195785

[82] S. A. Shamma, "Speech processing in the auditory system. i: The representation of speech sounds in the responses of the auditory nerve," *The Journal of the Acoustical Society of America*, vol. 78, pp. 1612–1621, 1985. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/4067077/

[83] ——, "Speech processing in the auditory system. ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *The Journal of the Acoustical Society of America*, vol. 78, pp. 1622–1632, 1985. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/3840813/

[84] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 421–435, 1994.

[85] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, pp. 824–839, 1992.

[86] P. Ru, "Perception-based multi-resolution auditory processing of acoustic signal," Ph.D. dissertation, University of Maryland, College Park, MD, 2000.

[87] S. A. Shamma, "Spatial and temporal processing in central auditory networks," in *Methods in Neuronal Modeling*, C. Koch and I. Segev, Eds. MIT Press, 1989, pp. 247–289.

[88] R. Lyon and S. Shamma, *Auditory Representations of Timbre and Pitch*. Springer, New York, NY, 1996, pp. 221–270. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4612-4070-9_6

[89] J. R. Cohen, "Application of an auditory model to speech recognition," *The Journal of the Acoustical Society of America*, vol. 85, pp. 2623–2629, 1989. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/2526170/

[90] "matlab implementations of all the above auditory spectrogram computations are available for download at https://isr.umd.edu/labs/nsl/software.htm."

[91] J. Laudanski, B. Torben-Nielsen, I. Segev, S. Shamma, H. Markram, C. Colbert, and M. Migliore, "Spatially distributed dendritic resonance selectively filters synaptic input," *PLoS Computational Biology*, vol. 10, p. e1003775, 8 2014. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1003775

[92] J. C. R. Licklider, "A duplex theory of pitch perception," *The Journal of the Acoustical Society of America*, vol. 23, pp. 147–147, 1 1951. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1917296

[93] G. Langner, "Periodicity coding in the auditory system," *Hearing Research*, vol. 60, pp. 115–142, 7 1992.

[94] G. Langner and C. E. Schreiner, "Periodicity coding in the inferior colliculus of the cat. i. neuronal mechanisms," *Journal of neurophysiology*, vol. 60, pp. 1799–1822, 1988. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/3236052/

[95] A. Rees and A. R. Møller, "Stimulus properties influencing the responses of inferior colliculus neurons to amplitude-modulated sounds." *Hearing research*, vol. 27, pp. 129–43, 1987. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3610842

[96] J. J. Eggermont, "Temporal modulation transfer functions for am and fm stimuli in cat auditory cortex. effects of carrier type, modulating waveform and intensity," *Hearing research*, vol. 74, pp. 51–66, 1994. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8040099/

[97] G. Stuart, N. Spruston, B. Sakmann, and M. Häusser, "Action potential initiation and backpropagation in neurons of the mammalian cns," *Trends in neurosciences*, vol. 20, pp. 125–131, 3 1997. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/9061867/

[98] S. Song, K. D. Miller, and L. F. Abbott, "Competitive hebbian learning through spike-timing-dependent synaptic plasticity," *Nature Neuroscience*, vol. 3, pp. 919–926, 9 2000. [Online]. Available: https://www.nature.com/articles/nn0900_919

[99] R. Kempter, W. Gerstner, and J. L. van Hemmen, "Hebbian learning and spiking neurons," *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 59, pp. 4498–4514, 4 1999. [Online]. Available: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.59.4498

[100] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press, 1949.

[101] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps," *Science (New York, N.Y.)*, vol. 275, pp. 213–215, 1 1997. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8985014/

[102] R. S. Heffner, G. Koay, and H. E. Heffner, "Audiograms of five species of rodents: Implications for the evolution of hearing and the perception of pitch," *Hearing Research*, vol. 157, pp. 138–152, 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11470193/

[103] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design," *Journal of Computational Neuroscience*, vol. 9, pp. 85–111, 2000.

[104] S. Norman-Haignere, N. Kanwisher, and J. H. McDermott, "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*, vol. 33, p. 19451, 12 2013. [Online]. Available: /pmc/articles/PMC3916670//pmc/articles/PMC3916670/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3916670/

[105] F. A. Bilsen, "Pronounced binaural pitch phenomenon," *The Journal of the Acoustical Society of America*, vol. 59, p. 467, 8 1998. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.380892

[106] E. M. Cramer, W. H. Huggins, E. . t, and M. Cramer, "Creation of pitch through binaural interaction," *The Journal of the Acoustical Society of America*, vol. 30, p. 413, 6 2005. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.1909628

[107] J. F. Culling, "The existence region of huggins' pitch," *Hearing Research*, vol. 127, pp. 143–148, 1 1999.

[108] M. A. Klein and W. M. Hartmann, "Binaural edge pitch," *The Journal of the Acoustical Society of America*, vol. 70, pp. 51–61, 1981. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/7264072/

[109] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *Citation: The Journal of the Acoustical Society of America*, vol. 35, p. 1206, 1963. [Online]. Available: https://doi.org/10.1121/1.1918675

[110] J. Adams, "Projections from octopus cells of the posteroventral cochlear nucleus to the ventral nucleus of the lateral lemniscus in cat and human," *Auditory Neuroscience*, pp. 335–350, 1997.

[111] G. D. Pollak, R. M. Burger, and A. Klug, "Dissecting the circuitry of the auditory system," *Trends in Neurosciences*, vol. 26, pp. 33–39, 1 2003.

[112] C. Pantev, T. Elbert, B. Ross, C. Eulitz, and E. Terhardt, "Binaural fusion and the representation of virtual pitch in the human auditory cortex," *Hearing Research*, vol. 100, pp. 164–170, 10 1996.

[113] M. S. Osmanski, X. Song, and X. Wang, "The role of harmonic resolvability in pitch perception in a vocal nonhuman primate, the common marmoset (callithrix jacchus)," *Journal of Neuroscience*, vol. 33, pp. 9161–9168, 5 2013. [Online]. Available: https://www.jneurosci.org/content/33/21/9161https://www.jneurosci.org/content/33/21/9161.abstract

[114] W. S. Rhode, D. Oertel, and P. H. Smith, "Physiological response properties of cells labeled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus," *J Comp Neurol*, vol. 213, pp. 448–463, 1983. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/6300200/

[115] P. H. Smith and W. S. Rhode, "Structural and functional properties distinguish two types of multipolar cells in the ventral cochlear nucleus," *The Journal of comparative neurology*, vol. 282, pp. 595–616, 1989. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/2723154/

[116] I. M. Winter, A. R. Palmer, L. Wiegrebe, and R. D. Patterson, "Temporal coding of the pitch of complex sounds by presumed multipolar cells in the ventral cochlear nucleus," *Speech Communication*, vol. 41, pp. 135–149, 8 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167639302000985

[117] I. M. Winter and A. R. Palmer, "Level dependence of cochlear nucleus onset unit responses and facilitation by second tones or broadband noise," *Journal of neurophysiology*, vol. 73, pp. 141–159, 1995. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/7714560/

[118] D. Oertel, R. Bal, S. M. Gardner, P. H. Smith, and P. X. Joris, "Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus," *PNAS*, vol. 97, 2000. [Online]. Available: http://www.pnas.org/content/97/22/11773.full.pdf

[119] N. L. Golding, D. Robertson, and D. Oertel, "Recordings from slices indicate that octopus cells of the cochlear nucleus detect coincident firing of auditory nerve fibers with temporal precision," *The Journal of Neuroscience*, vol. 15, pp. 313–3153, 1995. [Online]. Available: http://www.jneurosci.org/content/jneuro/15/4/3138.full.pdf

[120] G. D. Langner and C. Benson, "The neural code of pitch and harmony," *The Neural Code of Pitch and Harmony*, pp. 1–227, 1 2015. [Online]. Available: https://www.cambridge.org/core/books/neural-code-of-pitch-and-harmony/F6517CCE7D9F0E728D2689F22216F877

[121] D. Oertel, S. Wright, X. J. Cao, M. Ferragamo, and R. Bal, "The multiple functions of t stellate/multipolar/chopper cells in the ventral cochlear nucleus," *Hearing research*, vol. 276, pp. 61–69, 6 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21056098/

[122] M. Sachs and C. Blackburn, "Processing of complex sounds in the cochlear nucleus," in *Neurophysiology of Hearing: The central auditory system*, R. Altschular, R. Bobbin, and B. Clopton, Eds. Raven Press, 1991.

[123] F. H. Willard and G. F. Martin, "The auditory brainstem nuclei and some of their projections to the inferior colliculus in the north american opossum," *Neuroscience*, vol. 10, pp. 1203–1232, 1983. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/6664491/

[124] M. S. Malmierca, T. B. Leergaard, V. M. Bajo, J. G. Bjaalie, and M. A. Merchán, "Anatomic evidence of a three-dimensional mosaic pattern of tonotopic organization in the ventral complex of the lateral lemniscus in cat," *The Journal of Neuroscience*, vol. 18, p. 10603, 12 1998. [Online]. Available: /pmc/articles/PMC6793352//pmc/articles/PMC6793352/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6793352/

[125] S. H. Wu, "Physiological properties of neurons in the ventral nucleus of the lateral lemniscus of the rat: intrinsic membrane properties and synaptic responses," *Journal of neurophysiology*, vol. 81, pp. 2862–2874, 1999. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10368403/

[126] G. Langner, S. Braun, C. Simonis, C. Benso, and N. Cant, "New evidence for a pitch helix in the ventral nucleus of the lateral lemniscus in the gerbil," *Proceedings of the Associates for Research in Otolaryngology Abstracts*, p. 771, 2 2006.

[127] R. Batra, "Responses of neurons in the ventral nucleus of the lateral lemniscus to sinusoidally amplitude modulated tones," *Journal of Neurophysiology*, vol. 96, pp. 2388–2398, 2006.

[128] C. Micheyl and A. J. Oxenham, "Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings," *Hearing Research*, vol. 266, pp. 36–51, 7 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19788920http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2885481http://linkinghub.elsevier.com/retrieve/pii/S0378595509002366

[129] V. M. Richards, E. M. Carreira, and Y. Shen, "Toward an objective measure for a "stream segregation" task," *The Journal of the Acoustical Society of America*, vol. 131, p. EL8, 12 2011. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.3664107

[130] W. T. Newsome and E. B. Paré, "A selective impairment of motion perception following lesions of the middle temporal visual area (mt)," *Journal of Neuroscience*, vol. 8, pp. 2201–2211, 6 1988. [Online]. Available: https://www.jneurosci.org/content/8/6/2201https://www.jneurosci.org/content/8/6/2201.abstract

[131] S. Teki, M. Chait, S. Kumar, S. Shamma, and T. D. Griffiths, "Segregation of complex acoustic scenes based on temporal coherence," *eLife*, vol. 2013, 7 2013.

[132] E. Holmes and T. D. Griffiths, "'normal' hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception," *Scientific Reports 2019 9:1*, vol. 9, pp. 1–11, 11 2019. [Online]. Available: https://www.nature.com/articles/s41598-019-53353-5

[133] J. A. O'Sullivan, S. A. Shamma, and E. C. Lalor, "Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening." *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 35, pp. 7256–63, 5 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25948273

[134] F. Schneider, F. Balezeau, C. Distler, Y. Kikuchi, J. van Kempen, A. Gieselmann, C. I. Petkov, A. Thiele, and T. D. Griffiths, "Neuronal figure-ground responses in primate primary auditory cortex," *Cell Reports*, vol. 35, p. 109242, 6 2021.

[135] B. G. Shinn-Cunningham and V. Best, "Selective attention in normal and impaired hearing," *Trends in Amplification*, vol. 12, p. 283, 2008. [Online]. Available: /pmc/articles/PMC2700845//pmc/articles/PMC2700845/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2700845/

[136] M. A. Johns, R. C. Calloway, I. Phillips, V. P. Karuzis, K. Dutta, E. Smith, S. A. Shamma, M. J. Goupell, and S. E. Kuchinsky, "Performance on stochastic figure-ground perception varies with individual differences in speech-in-noise recognition and working memory capacity,," *Journal of the Acoustical Society of America (accepted)*, 2022.

[137] R. C. DeCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, pp. 1439–1443, 5 1998. [Online]. Available: https://www-science-org.proxy-um.researchport.umd.edu/doi/10.1126/science.280.5368.1439

[138] D. T. Blake and M. M. Merzenich, "Changes of ai receptive fields with sound density," *Journal of Neurophysiology*, vol. 88, pp. 3409–3420, 12 2002. [Online]. Available: https://journals.physiology.org/doi/10.1152/jn.00233.2002

[139] S. Teki, M. Chait, S. Kumar, K. V. Kriegstein, and T. D. Griffiths, "Brain bases for auditory stimulus-driven figure–ground segregation," *Journal of Neuroscience*, vol. 31, pp. 164–171, 1 2011. [Online]. Available: https://www.jneurosci.org/content/31/1/164https://www.jneurosci.org/content/31/1/164.abstract

[140] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLoS biology*, vol. 7, 6 2009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19529760/

[141] S. Teki, N. Barascud, S. Picard, C. Payne, T. D. Griffiths, and M. Chait, "Neural correlates of auditory figure-ground segregation based on temporal coherence," *Cerebral Cortex (New York, NY)*, vol. 26, p. 3669, 9 2016. [Online]. Available: /pmc/articles/PMC5004755//pmc/articles/PMC5004755/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5004755/

[142] C. Micheyl, H. Kreft, S. Shamma, and A. J. Oxenham, "Temporal coherence versus harmonicity in auditory stream formation," *The Journal of the Acoustical Society of America*, vol. 133, p. EL188, 2013. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/133/3/10.1121/1.4789866

[143] M. P. Layng and P. N. Chase, "Stimulus-stimulus pairing, matching-to-sample testing, and emergent relations," *The Psychological Record*, vol. 51, pp. 605–605, 9 2001.

[144] S. Atiani, S. V. David, D. Elgueda, M. Locastro, S. Radtke-Schuller, S. A. Shamma, and J. B. Fritz, "Emergent selectivity for task-relevant stimuli in higher-order auditory cortex," *Neuron*, vol. 82, pp. 486–499, 4 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24742467http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4048815http://linkinghub.elsevier.com/retrieve/pii/S0896627314001603

[145] F. Samson, K. L. Hyde, A. Bertone, I. Soulières, A. Mendrek, P. Ahad, L. Mottron, and T. A. Zeffiro, "Atypical processing of auditory temporal complexity in autistics," *Neuropsychologia*, vol. 49, pp. 546–555, 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21192958/

[146] D. Elgueda, D. Duque, S. Radtke-Schuller, P. Yin, S. V. David, S. A. Shamma, and J. B. Fritz, "State-dependent encoding of sound and behavioral meaning in a tertiary region of the ferret auditory cortex," *Nature neuroscience*, vol. 22, pp. 447–459, 3 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30692690/

[147] J. O'Sullivan, J. Herrero, E. Smith, C. Schevon, G. M. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, "Hierarchical encoding of attended auditory objects in multi-talker speech perception," *Neuron*, vol. 104, pp. 1195–1209.e3, 12 2019.