Roles of integration and stimulus history in the representation of sound by the auditory cortex of ferrets.

By

Mateo López Espejo

A DISSERTATION

Presented to the Neuroscience Graduate Program,

and Oregon Health & Science University

School of Medicine

In partial fulfillment of

the requirements for the degree of:

Doctor of Philosophy

December 2022

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD Dissertation of

Mateo López Espejo

Has been approved:

Advisor, Stephen V. David

Member and Chair, Haining Zhong

Member, Skyler Jackman

Member, Vincent Costa

External member, Santiago Jaramillo

Contents

Chapter 1. Introduction

The general purpose of the sensory brain

Our senses are continuously bombarded by an avalanche of stimuli. The sensory epithelia, and the downstream regions of the brain a few synapses away from them, have the task of finding and representing the parts of this sensorium which are relevant for behavior and survival.

This is a difficult task as the raw input of sensory systems, i.e., the activity of the sensory neurons (rod and cones in vision, hair cells in hearing) is very high dimensional where each sensory neuron corresponds to one dimension. Furthermore, the activity of neighboring sensory neurons is highly correlated and therefore carries redundant information, and there are temporal correlations as sensory experience tends to change smoothly over time.

The sensory brain maps this high dimensional highly correlated raw sensory input to a lower-dimensional, less redundant representation (Chechik et al., 2006) by projecting the information onto the activity of fewer and less correlated neurons. This is ultimately useful for behavior as it leads to the explicit representation of relevant abstractions. For example, as you are cast away in a distant tropical shore, the roar of a tiger and the crashing of waves will cause activity in all hair cells of both cochleas, however different downstream groups of neurons fire in these two cases, allowing decision and motor centers to readily identify the danger and enact an escape upon receiving input from the roaring neurons.

This mapping is not trivial, it is a complex nonlinear function. In the case of hearing, these mapping functions include the integration over diverse time scales of spectral information of sound. For humans such mapping transforms pressure waves on the cochlea into speech, which carries complex information (semantics, prosody) across multiple time scales (phonemes, syllables, words, phrases). An example of such integration in the scale of phrase semantics is that of homophones, where the meaning of a word *lies* in its context. Without the proper context,

truths might become *lies.* This complex mapping is performed at multiple steps along the different centers of the ascending auditory pathway(Escabí & Read, 2003; Norman-Haignere et al., 2022).

Hierarchical processing of auditory information

Anatomical hierarchy of the ascending auditory pathway

Sound is first sensed by the cochlea, which performs a spectral decomposition of the waveforms, a logarithmic compression of their amplitude and conversion into precisely timed electrical and chemical signals. Low to high frequencies, are separated and represented along the longitude of the cochlea, from the proximal to distal end relative to the middle ear (Fettiplace, 2020). This frequency ordered response, called cochleotopy or tonotopy is orderly transmitted forward through the cochlear nerve.

These signals are relayed to the cochlear nucleus and the olivary complex located in the brainstem, to later converge towards the Inferior Colliculus (IC) in the midbrain, and then the Medial Geniculate Body (MGB) in the thalamus. The thalamus then sends extensive and branching projections to the auditory cortex, reaching mostly the layer IV of the primary auditory cortex (A1), but with some projections to other cortical fields and layers (Huang & Winer, 2000). Through these relay, sound signals go from 3500 inner hair cells in each human cochlea to at least ten thousand time more neurons in A1 (DeWeese et al., 2005).

The auditory cortex is subdivided into regions based on the pattern of thalamic and cortical projections. After the lemniscal thalamic input reaches A1 (Bizley et al., 2005; Wallace et al., 1997), it is then transmitted to surrounding secondary regions, known as the belt and para-belt areas. These connections are both direct (cortico-cortical) or through a cortico-thalamic loop (Winer et al., 2005). These secondary regions are thought to further process information coming from A1. This hierarchical division of the cortex is further supported by response differences with secondary regions showing longer response latency, and broader tuning (Atiani et al., 2014; Bizley et al., 2005; Norman-Haignere et al., 2022).

Stepwise emergence of abstractions

This anatomical architecture of transmission relays has implications for the processing of information, it enables a stepwise processing of the auditory input. Early relays have access to more complete auditory information. This is the case in the cochlear nucleus and the olivary complex, where the activity of neurons faithfully follows the sound power at different frequencies, and in the case of low frequency, even the peaks and valleys of the sound wave itself. Since most of the incoming auditory information is represented, we can say that behaviorally important features of sound are also present, albeit in an implicit manner. As we have mentioned, this raw information is transformed and mapped into useful representations. An example of this transformation happens early in the auditory pathway at the medial superior olive (MSO). Neurons in this area have access to low latency wave phase information coming from both ears, and function as coincidence detectors that can represent the time difference of sound arrival between ears (Ashida & Carr, 2011). We can say that the interaural time difference information that was implicitly represented in the cochleas is then explicitly represented (abstracted) in the MSO. This explicit representation can then be used to infer the azimuthal position of sound sources. However, in the process the MSO loses some information that was present in the cochleas. This process of transforming implicit to explicit representations happens at every step of the ascending auditory pathway.

In the next auditory center, the inferior colliculus, auditory information is integrated with — and modulated by — other sensory systems. For example, information from eye position and visual input, alongside interaural time differences can be used to infer the location of a sound source (Gruters & Groh, 2012). Auditory information then passes through the thalamus, which filters and gates information to and between regions of the auditory cortex (Huang & Winer, 2000). Neurons in the MGB and the primary auditory cortex respond similarly to sound, however, the transformations of auditory information that happen at the thalamocortical interface are numerous and complex.

Inheritance, construction, and ensemble

The thalamocortical interface computation can be summarized in 3 heuristic rules (Miller et al., 2001): (i) 'Inheritance', where the tuning of cortical neurons is equal to that of their presynaptic thalamic partners (ii) 'Construction', where the tuning of multiple thalamic neurons is summed, conferring the cortical neuron a much broader tuning (iii) 'Ensemble', where only the common part of the multiple thalamic receptive fields is present in the cortical tuning. The mechanism of this ensemble or intersection might include lateral blanket inhibition suppressing the response to all but the common part of the stimulus (Winer et al., 2005).

These three mechanisms describe computations performed on the spatial or spectral dimensions of sensory stimuli. Vision offers an example of "construction" where the union of center surround fields generate simple fields (Hubel & Wiesel, 1962). Hearing, on the other hand, is predominantly represented by "inheritance", where the tuning of cortical neurons match that of their thalamic presynaptic partners (Miller et al., 2001).

This inheritance would seem like a superfluous step: the transmission of unchanged information. However, while tuning is inherited, differences in the temporal properties of neurons in A1 emerge, where sound evoked responses come with an extended latency due to the extra synaptic steps, but also an increased duration in the responses, which corresponds to an increased temporal integration window. This suggests that the thalamocortical interface computation is predominantly temporal. However, spectral computations like inhibition by sounds outside the preferred frequency, help sharpen the tuning to the preferred frequency in the auditory cortex (Kato et al., 2017; Lakunina et al., 2020), and paradoxically might contribute to the apparent simple inheritance of tuning from the thalamus.

This increase in the temporal integration it not only present at the thalamocortical interface, but along the entirety of the ascending auditory pathway (Asokan et al., 2021; Escabí & Read, 2003) and between primary and secondary regions of the auditory cortex (Atiani et al., 2014; Norman-Haignere et al., 2022).

While these three scenarios were proposed for the thalamocortical interface, variations of them might appear on other brain regions. Furthermore, they only consider bottom-up information transmission, which ignores the top-down feedback signals, and the local computations performed at the IC, MGB and A1. It also obviates some of the more complex nonlinear interactions between inputs to a cortical neuron, where the operation performed is not a simple union or intersection. The contribution of these different sources of information and tuning are the subjects of ongoing research.

Phase and rate codes

A main consequence of these temporal computations, characterized by an increase in the integration window and the duration of neuronal responses is the transformation of implicit *phase codes* into explicit *rate codes*, two strategies to represent temporal modulation of sound.

Phase codes are found in early stages of the auditory pathway, where neuronal activity is said to be phase locked to sound, i.e., it faithfully follows the peaks and valleys of an amplitude modulated sound. The precise timing of phase codes is necessary for some computations. For example, inferring the position of sound sources from binaural queues (Ashida & Carr, 2011). Precise neuronal timing is enabled by special characteristics of these early auditory neurons, that ensure fast and reliable synaptic transmission, with low latency, jitter and temporal summation, i.e., short integration windows (Trussell, 1999).

Rate codes appear at later stages in the auditory pathway. As the temporal integration accumulates over serial synapses, neurons become unable to resolve amplitude modulation of frequencies (period) faster than their integration window. Instead, these neurons modulate their firing rate as a function of the amplitude modulation rate of sound (T. Lu et al., 2001). Furthermore, the feedforward architecture and expansions motifs of the leminiscal auditory pathway favors rate code transmission (Barral et al., 2019).

Single neuron representations

We have referred to the spectral preference or tuning of a neuron, and to the temporal properties of its response, latency, and duration. This, however, is an oversimplification of the response characteristics of neurons in the auditory cortex, which are tuned to precise spectrotemporal characteristics of sound. For example, it's not uncommon to find neurons which respond to the onset of a sound, but quickly stop their firing even when the sound continues. Is this neuron activated or suppressed by this frequency? What is the temporal window of these activation and suppression, and how are these two integrated together?

The STRF

To answer these questions requires modeling the response properties of the neuron. However, the answers given by a model will be contingent on the assumptions and biases of the model itself. It's therefore wise to use models with few of these assumptions, such as reverse correlation, where the occurrence of spikes or changes in firing rate are correlated with the power present at different spectral bands and time lags of the sound. These combinations of frequency and time (the two orthogonal dimensions of sound) can be represented as a spectrogram.

The reverse correlation of neuronal activity to these sound spectrograms yields the spectrotemporal receptive field (STRF) (Aertsen & Johannesma, 1981). Despite the recent developments in machine learning and AI, the STRFs remain an unbiased and powerful description of the tuning of auditory neurons, and how this tuning is transformed along the auditory pathway. Furthermore, STRFs enable the inference of a neuron tuning with more efficient stimuli like temporally orthogonal ripple combinations (TORC) (Klein et al., 2000), but also with natural sounds (Theunissen et al., 2001).

STRF changes along the auditory pathway

STRF of neurons along the auditory pathway confirms that the main computations performed early on are on the time domain. STRFs from IC, thalamus and A1 share a similarly narrow

spectral tuning (Miller et al., 2002; Sen et al., 2001), whereas the temporal windows of receptive field increase from responses as fast as 10ms in the IC to past 100ms in A1 (Escabí & Read, 2003; Sen et al., 2001).

Once in the auditory cortex, more spectral-related computations appear and STRFs become higher dimensional, i.e., described by multiple spectral and temporal filters (Atencio et al., 2012). In the secondary regions of the cortex, the frequency tuning becomes broader, exemplified by neurons responding only to broadband noise, but not pure tones (Bizley et al., 2005).

Alongside this cortical spectral integration, neurons with more complicated and specific tunings emerge. An example is that of neurons tuned to the vocalizations of conspecifics, but unresponsive to tones, TORCs, and other synthetic sounds with equal spectral content (Montes-Lourido et al., 2021; Rauschecker et al., 1995; Theunissen et al., 2000). Capturing these more complex responses with STRFs or more complex models, will help to understand why neurons respond to certain sound properties only during specific contexts for example, as part of a vocalization. Furthermore, if the computations behind receptive fields are elucidated, they can guide the search for their underlying physiological substrate. However, modeling these receptive fields is not a trivial task.

Temporal and linear limitations of the STRF

The classic STRF cannot capture more complex tunings. This is due to the linear nature of the STRF, which cannot fit the nonlinear interactions which precisely give rise to the specificity and complexity of the tuning of some of these later cortical neurons. Nonlinearities are accumulated at every neuron along the auditory pathway, with their firing threshold, membrane time constants, synaptic adaptation, integration across multiple synaptic inputs and time windows, etc.

A simple example of nonlinearity is the minimum and maximum responses of a neuron associated with zero spikes (rectification), and the neuron maximum firing rate (saturation). For this nonlinearity, an STRF can be extended with a sigmoid function, which captures this

rectification and saturation of firing rate (Thorson et al., 2015). This is known as a linear non-linear model (LN-model). The question turns into finding the right nonlinearities to extend and improve the STRF.

Empirical observations have shown the temporal limitations of STRFs, which fail to describe the effects of past sounds, more than 150 ms ago, on neuronal response (Atiani et al., 2014). However, cortical neurons' response and integration windows extend past the STRF estimations, as evidenced by extended offset responses (Schinkel-Bielefeld et al., 2012), and the influence of recent sounds on the response to an ongoing stimulus (Angeloni & Geffen, 2018; Asari & Zador, 2009).

This STRF temporal limitation is likely caused by the nonlinear nature of the long duration influence of sound on neuronal activity. The likely underlying mechanisms are synaptic plasticity, the intrinsic properties of neurons (Dean et al., 2008; Whitmire & Stanley, 2016), and local neuronal population dynamics (Dean et al., 2005). It is worth exploring the known physiology behind these mechanisms to properly model the appropriate extensions to the STRF to capture them.

Mechanisms of temporal processing

Adaptation

As mentioned, adaptation is one potential source of nonlinearities, however, adaptation plays other equally important roles in sound representation.

Neurons efficiently respond to and relay the broad range of activity arriving to them. This is achieved by adjusting their transfer function (input to firing rate) to the statistics of the input. For the auditory system this means encompassing a broad dynamic range that goes from the quiet rustle of leaves in the jungle, to the roar of a tiger — which is orders of magnitude louder. Adaptations to input statistics are not limited to sensory neurons, they appear on most neurons in the brain as they adapt to all their synaptic inputs (Barlow, 2012; Beyeler et al., 2019).

Adaptation is supported by synaptic mechanisms like the changes in the release probability of synaptic vesicles, which are mediated by the interplay of activity-dependent Ca2+ concentration, the size of the readily releasable pool, and the dynamics of the Ca sensitive SNARE protein complex (for an exhaustive review see (Jackman & Regehr, 2017)). Furthermore, the neuron has intrinsic passive and active electrical properties, which actively modulate its firing threshold (Silver, 2010).

Other mechanisms of synaptic plasticity, like long term potentiation, will not be discussed here as their durations go beyond the time scale of hundreds of milliseconds to seconds, associated with the representation of ongoing stimuli in the current soundscape context.

The two directions of synaptic adaptation, facilitation and depression are thought to play distinct roles as temporal filters. Facilitation enables the transmission of sustained burst of activity, acting as a high pass filter, while depression transmits changes in firing rate with, acting as a low pass filter (Jackman & Regehr, 2017).

Besides these filtering properties, the state of depression or facilitation of a synapse holds information about its recent levels of activity. This information is not explicitly represented in the firing rate of the neuron, but instead remains “hidden” until revealed by a new bout of activity.

The memory held by adaptation is limited by the time the neurons take to relax back to their basal states. As mentioned before, these time constants are consequence of the complex interplay of many synaptic proteins. The complexity of this system is further compounded by the number of synapses and neurons and neuron types that are involved in the local circuitry integrating and representing sound stimuli.

To overcome and model the multitude of this synaptic and neuronal complexity we can make use of some simpler approximation. This has been done by describing adaptation as a change in synaptic strength (either increase or decrease) with each subsequent action potential, and a return to a baseline strength following an exponential decay (Tsodyks et al., 1998). This simplified model describes adaptation in two parameters corresponding to synaptic availability and rate of

recovery. Furthermore, this equation can be used multiple times in parallel to account for the different time constant and integration windows that might be involved and can be readily implemented for firing rates making it compatible to STRFs inferred from extracellular recordings.

Role of adaptation on the representations of sounds

While adaptation happens on a timescale of hundreds of milliseconds, it might be associated to representation of sound features on much longer time scale, for example those associated to adaptation to background environmental sounds like the rain in the jungle, but that does not reduce the response to a salient stimulus, like the distant roar of a tiger, or the distinct *repeating pattern* of splashing sounds made by an animal running through the wet jungle.

The extent to which the auditory brain can filter out background stimuli, while readily representing salient ones has been extensively studied in human cognition through EEG and MEG measurements. A standard measurement of stimulus salience is the mismatch negativity (MMN), in which event-related potentials are characterized by a reduced response to repeated (standard) stimuli and an enhanced response to unexpected (oddball) sounds (Ulanovsky et al., 2003). The characteristics that define standard and oddball range from straightforward sound dimensions like frequency, to more complex and abstract dimensions, like the perceived gender of a speaker, i.e., the sudden voice of a woman pops out from repeated male speech, independent of the content of what is said (Casado & Brunellière, 2016).

An effect related to the MMN occurs during pattern detection. Listeners are presented with a random sequence of tones (standard) which at some point transitions to a regular, repeating sequence (oddball). In this case, they can readily detect when the pattern starts repeating and regularity emerges. Furthermore, the neural correlates of this percept are observed even when subjects are performing an unrelated visual task. Thus this pattern detection is pre-attentive, and likely supported by a bottom-up integration mechanism (Barascud et al., 2016).

Stimulus specific adaptation

How neuronal adaptation can give rise to these macroscopic observations have been studied through the lens of stimulus-specific adaptation (SSA) (Ulanovsky et al., 2003). Similar to MMN, neurons tend to decrease their response to repetitive (standard) sounds, without changing — and sometimes increasing — their response to unexpected (oddball) stimulus. Some SSA is observed starting from the IC. However, it becomes more pronounced in A1 (Carbajal & Malmierca, 2018), and occurs for different dimensions of sound like frequency and inter stimulus intervals. Given its similarities, SSA is thought to be an underlying mechanism of MMN.

SSA is partly mediated by bottom-up processing, where different synaptic inputs receive distinct narrowly tuned inputs, i.e., through the “construction” process characteristic of the thalamocortical interface. However, other mechanisms, such as recurrent top-down influences and local circuit dynamics, also contribute to SSA. Within local circuitry, the role of inhibitory interneurons (IN) is of note, as they mediate a refinement and amplification and shaping of SSA (Natan et al., 2015; Yarden et al., 2022).

Cortical column circuitry

The mechanism of how distinct neuronal subtypes contribute to SSA, context integration and other temporal computations must be explained from the perspective of their general role in local circuitry. For A1, this circuitry is instantiated as the archetypal cortical column (Mountcastle, 1997), characterized by its 6-layer organization, its presence across all mammalian cortex, and the specific computations it supports.

First, thalamic input reaches mostly neurons in the granular layers (L4). Granular neurons therefore show simple receptive fields emerging from inheritance, construction, and ensemble, evidenced by simple, and single-whisker neurons in the L4 of visual and barrel cortex respectively. Information propagates to the infra-granular (L5, L6) and supra-granular (L1, L2/3) layers. In the latter, additional cortico-cortical projections arrive. This generates more complex receptive fields in infra- and supra-granular layers, evidenced by the enrichment of hypercomplex and multi-

whisker neurons in the visual and barrel cortex respectively (Brumberg et al., 1999; Hubel & Wiesel, 1962; Martinez & Alonso, 2003).

Although tuning to sound frequency is distributed tonotopically, neurons in auditory cortex do integrate information across frequency channels as well (Linden & Schreiner, 2003; A. K. Moore & Wehr, 2013; Tischbirek et al., 2019).There is evidence of frequency integration in the form of sideband suppression, which arrives through L2/3 cortico-cortical recurrent connections, from adjacent columns (Kato et al., 2017). It appears that the main systematic difference between auditory cortex layers is not content but computations, i.e., the complexity and nonlinearity of interactions between the different parts that compose the receptive field of neurons (Atencio et al., 2009). This supports the more subtle layer differences observed, like the reliable or sparse response to click trains in granular and supra-granular layers respectively (Sakata & Harris, 2009), the selective response to conspecific vocalizations in supra-granular layers (Montes-Lourido et al., 2021), and to the increased plasticity of STRFs in supra-granular layers during behavioral tasks (Francis et al., 2018).

Inhibitory interneuron's role in computations

Within the different cortical layers, different IN subtypes perform distinct roles in computation, which go beyond preventing runoff activity by matching and equilibrating excitation (E/I equilibrium) (Isaacson & Scanziani, 2011). Somatostatin (SST) expressing INs target the apical dendritic arbors exerting weak but facilitating, dendrite specific inhibition (Murayama et al., 2009). Parvalbumin (PV) expressing INs, exert strong, but quickly depressing suppression of somas, gating the response of pyramidal neurons with precise timing (Nocon et al., 2022). Vasoactive Intestinal Peptide (VIP) expressing INs target other INs and play a role in release from inhibition (Pi et al., 2013). For the auditory cortex, the particular temporal profiles of response and synaptic adaptation of different IN subtypes play a significant role in temporal computations (Seay et al., 2020).

Interestingly, besides this precise timing, INs also show responses extending past sound termination. Furthermore, these extended responses differ from those during sound in their sensitivity to recent sound history, the latter being more depressed relative to the former during conditions of sustained stimulation (i.e., Standard sound for the SSA oddball paradigm). This differential adaptation plays a role in representing the recent sound history, as PV neurons preferentially inhibit the response of pyramidal neurons under low adaptation (SSA deviants), while VIP reduce this inhibition under high adaptation (SSA standards) (Yarden et al., 2022). These two mechanisms constitute opposite roles which can then modulate the extent of adaptation to past neuronal history, therefore modulating the duration of the cortical integration window.

IN specializations and their associated temporal computations, a hallmark feature of auditory processing, are likely necessary for the emergence of tuning along more complex dimensions of sound stimuli, like conspecific calls selectivity, and ultimately speech.

Finally, despite the similar spectral tuning of neighboring neurons, IN and particularly SOM neurons play a role in refining the frequency tuning of neurons though inhibition elicited by sounds outside of the preferred frequency (Lakunina et al., 2020), which comes from differently-tuned neighboring columns though the wide ramifications of SOM neurons (Kato et al., 2017). In conjunction with the aforementioned temporal integration, these are likely mechanisms that identify the general statistics of noise, and subtract it from relevant signals, thus generating noise-invariant representations (Rabinowitz et al., 2013), and improving the detection of sound in noise (Lakunina et al., 2022).

We have treated the role of neurons as isolated parts of a circuit, representing segments of auditory features, like the different time scales associated to different cell types. However, this is an incomplete view, and to gain insight of the computations performed by circuits, we must consider the coordinated activity of populations of neurons that form these circuits. We must then turn our focus to the representation of sound by populations of neurons.

Distribution of representations in neuronal populations

The development and miniaturization of electronics allowed for the simultaneous recording of thousands of neurons, as is the case for Neuropixels (Steinmetz et al., 2021). Having access to the response of population of neurons inevitably leads to the question of how the labor of computation and representation are distributed within a population. If we think on temporal integration, and how responses to current sounds will be influenced by historic activity, we confront a combinatorial explosion, where the brain not only needs to represent a sound but also all its variations when embedded in diverse historical contexts. The value of distributed population representations becomes evident with the richness of the soundscape.

Efficient sparse code

Historically, population sensory coding has been studied from the perspective of efficient sparse coding (Lewicki, 2002). Efficient insofar as the representations in the brain tend to match the statistics of the sensory inputs, i.e., not all sounds are equally likely to occur in nature, and therefore evolutionary pressure devoted more neural resource to represent naturally occurring sounds. Sparse insofar as how distinct categories and features of sound are represented by different subsets of neurons, i.e., distinctly and specifically tuned neurons.

Dense and local codes: capacity, robustness, and readout

Sparse codes lay on a continuum between two extreme strategies, a local code, and a dense code. These two extreme code modalities have advantages and disadvantages. In the local code, every neuron in the population is specialized in explicitly representing one abstraction, e.g., the concept of grandmother (Quiroga et al., 2005). The representation of frequency in the auditory nerve resembles this paradigm, in which distinct (groups of) neurons, or labeled lines, transmit specific frequency information. This strategy enables simple decoding and composition (as explained for the thalamocortical interface), e.g., a chord represented as the sum of different frequency neurons. However, a local code is limited in the amount of representation (one

representation per labeled line), and it's vulnerable to damage, where a representation would be lost if the labeled line carrying it was damaged. This coding scheme works for the auditory nerve where the sound feature being explicitly represented is frequency, which is relatively low dimensional compared to more complex abstractions emerging later.

As information is broadcasted to a greater number of neurons in the auditory cortex, and numerous and diverse abstractions emerge along the way, a local code is no longer suitable. Dense code presents an alternative, where all abstractions are represented by the weighted activity of all the neurons in the population. This distributed representation is robust to neuron death, and it has exponential storage capacity given by the equation M^N where M the different levels of activity (firing rates) of N the number of neurons. However, it's difficult to decode, and will not necessarily yield readily composable representations as described on the thalamocortical interface.

Sparse codes: A tradeoff for optimal coding

Sparse codes lie between dense and local codes, thus balancing capacity, composability, robustness, and decodability. Sparse codes are a set of components, akin to principal components in PCA, which are selected such that stimuli can be reconstructed as the weighted sum of as few (sparse) components as possible, i.e., as a weighted sum of all components, where most components are weighted zero. This seemingly arbitrary constraint yielded models with components that resemble the receptive fields of neurons across multiple brain areas. Since their introduction in the visual system (Olshausen & Field, 1996, 2004), sparse codes have also been adapted to the auditory field, and can capture multiple stages of representation, from gammatones describing the response of the auditory nerve (Lewicki, 2002), to STRFs in the IC (Carlson et al., 2012) up to the auditory cortex, now using deep learning models (Zhang et al., 2019). This success of sparse deep neural network models in capturing representations at different auditory regions, and of different levels of abstraction, supports the prevalence of sparseness as a general organizing principle for sensory representations. Similar sparseness might be present for the

representation of more abstract features emerging from nonlinearities across frequency and over increasing periods of time.

Temporal tiling

Populations codes can implement other strategies for the representation of temporal features, particularly those that might last longer than the integration window of individual neurons and cannot therefore be resolved by them. To overcome this limitation, neurons within a population coordinate their activity, tiling the duration of stimuli or task representations during behavior, as observed in A1 and the posterior parietal cortex (PPC), a region associated with visual and auditory integration into perceptual objects, objects-oriented action, and sound localization (Runyan et al., 2017). This temporal tiling can be thought of as a sparse code of the temporal location in the time interval associated to sensory experience or behavioral task.

Internal state effects on representation

So far, we have described how information flows bottom-up from the periphery (sensory epithelia) towards sensory cortices. However, a parallel descending pathway exists, which permits a top-down flow for these higher cognitive representations to influence sensory representations in the sensory cortices (Choi et al., 2018).

Studying these top-down influences of internal states and representations poses a challenge, since this state cannot be readily and precisely manipulated like sensory stimuli. Despite these limitations, some internal states like arousal and its associated noradrenergic and cholinergic activity can be inferred from pupil dilation (McGinley et al., 2015; Reimer et al., 2016), directly read form locomotion (Schneider et al., 2021), or can be enforced like attention during behavioral tasks.

Stimuli can influence these internal states, e.g., hearing the distant roar of a tiger leads to a state of fear, arousal, and heightened attention. This establishes a feedback loop in which perception alters internal states, which in turn modulate perception. Therefore, internal states

constitute another expression of the context of historical stimuli, which works on longer time scales, and is supported by top-down connections carrying auditory and multisensory information (Choi et al., 2018), and input from modulatory brain regions like the nucleus basalis (Froemke et al., 2007) . Consistent with this, models including arousal (pupil dilation) and behavioral aspects, like task engagement, can better capture variation in the sound evoked activity of IC and A1 neurons (Saderi et al., 2021), through the modulation of the overall excitability of cortical neurons (Schwartz et al., 2020).

Context as hidden state

We have drawn a complex picture of the sources of information, internal and external, and the transformation and temporal computations performed on different time scales by local circuits and recurrent loops between brain regions. Understanding the nuance of these computations, the complex interplay between local and brain wide connectivity and their different time scales, quickly becomes an intractable problem. The advent of deep learning and AI, which arose from inspiration coming from neuroscience, might now lend a helping hand back. Theoretical developments like liquid state machines, and state dependent computations (Buonomano & Maass, 2009), can help frame the problem. At their core these theories posit a system of interconnected spiking neurons with internal states defined by adaptation and synaptic plasticity, which receive a continuous stream of information. At any given point, this information propagates through the network, generating a pattern of spikes which in turn changes the adaptation state of neurons and synapses. These changes remain “hidden”, as they are not readily observable unlike spikes. This hidden state is the substrate that sustains memory of stimulus history and context. However, this hidden state can be read in the influence it will exert on the network response to new incoming stimuli. A simple example of this is the SSA which we have already treated, however, liquid state machines are universal function approximations, i.e., with enough neurons, and diversity in adaptation time constants, any arbitrary nonlinear function can be implemented (Maass &

Markram, 2004). As we have seen, these nonlinearities are necessary to generate more complex receptive fields and higher abstractions.

Liquid state machines are a general framework that encompasses diverse mechanisms like sustained activity, adaptation, and recurrent connections. This strength in generality can become a weakness: further refinements are needed to capture the specifics of their implementation by the brain. Care is required, as the hubris and ego of trying to capture all the nuance of the brain *In Silico* will lead to catastrophic results, as it has happened before with the Blue Brain Project (Hutton, n.d.).

Our contributions

In the following chapters we will discuss the specific mechanisms by which adaptation and population dynamics implement some of the more complex computations and representations of nonlinear sound features. In chapter 2 we demonstrate how adaptation occurs independently for different spectral inputs, i.e., a neuron might adapt to a low frequency sound but remain responsive to a higher frequency one. We show how this independent adaptation supports well characterized temporal computations, like SSA, and an overall increase in the variability in sound responses, which is required to capture the diversity of natural sounds. Response variability emerges from the diverse spectro-temporal preferences of neurons, and the combinations and nonlinear interactions between these preferences. In chapter 3 we dive deeper into the sources of diversity in sound responses by exploring the duration of the temporal integration window of neurons. We show how recent auditory input influences responses to ongoing sounds, such that information about recent and ongoing sounds coexist in the activity of populations of neurons forming a sparse code. We also show how this complex integration of recent and ongoing sound is supported by a combination of neuron adaptation, circuit connectivity, and population dynamics.

Chapter 2. Spectral tuning of adaptation supports coding of sensory context in auditory cortex

Adapted from (Lopez Espejo et al., 2019)

Abstract

Perception of vocalizations and other behaviorally relevant sounds requires integrating acoustic information over hundreds of milliseconds. Sound-evoked activity in auditory cortex typically has much shorter latency, but the acoustic context, i.e., sound history, can modulate sound evoked activity over longer periods. Contextual effects are attributed to modulatory phenomena, such as stimulus-specific adaptation and contrast gain control. However, an encoding model that links context to natural sound processing has yet to be established. We tested whether a model in which spectrally tuned inputs undergo adaptation mimicking short-term synaptic plasticity (STP) can account for contextual effects during natural sound processing. Single-unit activity was recorded from primary auditory cortex of awake ferrets during presentation of noise with natural temporal dynamics and fully natural sounds. Encoding properties were characterized by a standard linear-nonlinear spectro-temporal receptive field (LN) model and variants that incorporated STP-like adaptation. In the adapting models, STP was applied either globally across all input spectral channels or locally to subsets of channels. For most neurons, models incorporating local STP predicted neural activity as well or better than LN and global STP models. The strength of nonlinear adaptation varied across neurons. Within neurons, adaptation was generally stronger for spectral channels with excitatory than inhibitory gain. Neurons showing improved STP model performance also tended to undergo stimulus-specific adaptation, suggesting a common mechanism for these phenomena. When STP models were compared between passive and active behavior conditions, response gain often changed, but average STP parameters were stable. Thus, spectrally and temporally heterogeneous adaptation, subserved

by a mechanism with STP-like dynamics, may support representation of the complex spectro-temporal patterns that comprise natural sounds across wide-ranging sensory contexts.

Introduction

Vocalizations and other natural sounds are characterized by complex spectro-temporal patterns. Discriminating sounds like speech syllables requires integrating information about changes in their frequency content over many tens to hundreds of milliseconds (Binder et al., 2000; Huetz et al., 2011; Mesgarani, Cheung, et al., 2014). Models of sensory encoding for auditory neurons, such as the widely used linear-nonlinear spectro-temporal receptive field (LN model), seek to characterize sound coding generally. That is, they are designed to predict time-varying responses to any arbitrary stimulus, including natural sounds with complex spectro-temporal dynamics (Wu et al., 2006). When used to study auditory cortex, however, LN models typically measure tuning properties only with relatively short latencies (20-80 ms), which prevents them from encoding information about stimuli with longer latency (Atiani et al., 2014; deCharms et al., 1998; Depireux et al., 2001). It remains an open question how the auditory system integrates spectro-temporal information from natural stimuli over longer periods.

Classic LN models cannot account for integration over longer timescales, but studies of spectro-temporal context have shown that auditory-evoked activity can be modulated by stimuli occurring hundreds to thousands of milliseconds (Angeloni & Geffen, 2018; Asari & Zador, 2009; Klampfl et al., 2012) or even several minutes beforehand (K. Lu et al., 2018b; Yaron et al., 2012). These results have generally been interpreted in the context of pop-out effects for oddball stimuli (Carbajal & Malmierca, 2018; Natan et al., 2015; Ulanovsky et al., 2003; Yarden & Nelken, 2017) or gain control to normalize neural activity in the steady state (Dean et al., 2005; Lesica & Grothe, 2008b; Rabinowitz et al., 2011). Encoding models that incorporate recurrent gain control or nonlinear adaptation have been shown to provide better characterization of auditory-evoked

activity in the steady state, indicating that these properties of neurons may contribute to context-dependent coding on these longer timescales (S. V David & Shamma, 2013; Rabinowitz et al., 2012; Rahman et al., 2019; Williamson et al., 2016; Willmore et al., 2016; Yarden & Nelken, 2017). Some models have been shown to account for cortical responses to natural stimuli more accurately than the LN model (Harper et al., 2016; Kozlov & Gentner, 2016; Rahman et al., 2019; Willmore et al., 2016), and others have been proposed that have yet to be tested with natural stimuli (Ahrens et al., 2008; Atencio et al., 2008; Rabinowitz et al., 2012; Williamson et al., 2016). These findings suggest that an adaptation mechanism plays a central role in context-dependent coding, but there is no clear consensus on the essential components of a model that might replace the LN model as a standard across the field.

Short-term synaptic plasticity (STP) is a widely-observed phenomenon in the nervous system. Upon sustained stimulation, the efficacy of synapses is depressed or facilitated until stimulation ceases and synaptic resources are allowed to return to baseline (del Castillo & Katz, 1954; Tsodyks et al., 1998). Activity evoked by a sensory stimulus will engage synaptic plasticity across the auditory network, and the specific synapses that undergo plasticity will depend on the stimulus. Because the pattern of plasticity is stimulus-dependent, it could provide a latent code for sensory context that modulates responses to subsequent stimuli. Thus we hypothesized that nonlinear adaptation with STP-like properties may play a general role in auditory cortical processing. The precise mechanism producing nonlinear adaptation can take other forms than STP (e.g., feedforward inhibition, postsynaptic inhibition (Natan et al., 2015; Nelken, 2014)), but all these mechanisms support a simple and fundamentally similar algorithm for encoding spectro-temporal features. The focus of this study is whether functional properties of auditory neurons are impacted significantly by such a mechanism at the algorithmic level and, in particular, if this adaptation occurs independently across inputs with different sound frequency. Regardless of precise mechanism, a population of neurons with spectrally tuned adaptation may support a rich

code for information over the many hundreds of milliseconds required to discriminate spectro-temporally complex natural sounds (Buonomano & Maass, 2009; Fortune & Rose, 2001).

To test for spectrally tuned adaptation during auditory processing, we developed a vocalization-modulated noise stimulus in which two simultaneous noise bands are modulated by envelopes from independent natural vocalizations. The naturalistic dynamics of these stimuli produce a wide range of sensory contexts for probing neural activity. We presented these stimuli during single-unit recordings in primary auditory cortex (A1) of awake ferrets and compared the performance neural encoding models to test for STP-like effects (Thorson et al., 2015; Wu et al., 2006). We fit variants of the LN model in which inputs adapt either locally to one spectral band or globally across all channels. For many neurons, locally tuned adaptation provided a more accurate prediction of neural activity, supporting the idea of channel-specific adaptation. The strength and tuning of adaptation was heterogeneous across the A1 population, consistent with the idea that a diversity of spectrally tuned adaptation supports a rich basis for encoding complex natural sounds. We observed the same pattern of results for models fit to a library of fully natural sounds.

We also asked how changes in behavioral state, which can influence response gain and selectivity, affected nonlinear adaptation properties in A1 (Ding & Simon, 2012; Fritz et al., 2003; Kuchibhotla et al., 2016; Mesgarani & Chang, 2012; Niwa et al., 2012; Otazu et al., 2009; Schwartz & David, 2018). We compared model STP parameters between passive listening and during a behavior that required detecting a tone in a natural noise stream. While the gain of the neural response could fluctuate substantially with behavioral state, STP was largely stable across behavior conditions. This finding suggests that, unlike response gain, nonlinear adaptation properties are not influenced by behavioral state and may instead be critical for stable encoding of spectro-temporal sound features (Buonomano & Maass, 2009; Chance et al., 1998). Together, these findings demonstrate that during natural hearing, a simple, STP-like mechanism can explain many aspects of context-dependent sound coding. Moreover, these processes, typically

associated with steady-state adaptation to different contexts, such as SSA, can play a more dynamic role, continuously shaping the representation of spectro-temporally complex natural sounds.

Results

Encoding models reveal spectrally tuned adaptation in primary auditory cortex

This study characterized how primary auditory cortex (A1) integrates information from dynamic, naturalistic stimuli over frequency and time. Data were recorded from 200 single units in A1 of 5 passively listening ferrets during presentation of two band vocalization-modulated noise (Figure 1 A-B, (S. V David & Shamma, 2013; Lesica & Grothe, 2008a)). The stimulus contained complex natural temporal statistics but simple spectral properties. Thus it allowed an experimental focus on nonlinear temporal processing in the presence of multiple spectral features. Noise bands were one-quarter octave and modulated by different natural vocalization envelopes. Both bands were positioned so that they fell in the spectral receptive field of recorded neurons, as measured by briefly presented tones or noise bursts (Figure 1 B).

The dynamic vocalization-modulated noise often evoked reliable time-varying responses from A1 neurons, but the timecourse of this response varied substantially. Peri-stimulus time histogram (PSTH) responses computed from average repetitions of identical noise stimuli showed that responses could predominantly follow the envelope of one or both of stimulus bands (Figure 1 C). Thus, while all neurons included in the study were excited by isolated, narrowband stimuli in each frequency band, responses to stimuli presented in both bands simultaneously were complex and varied across neurons.

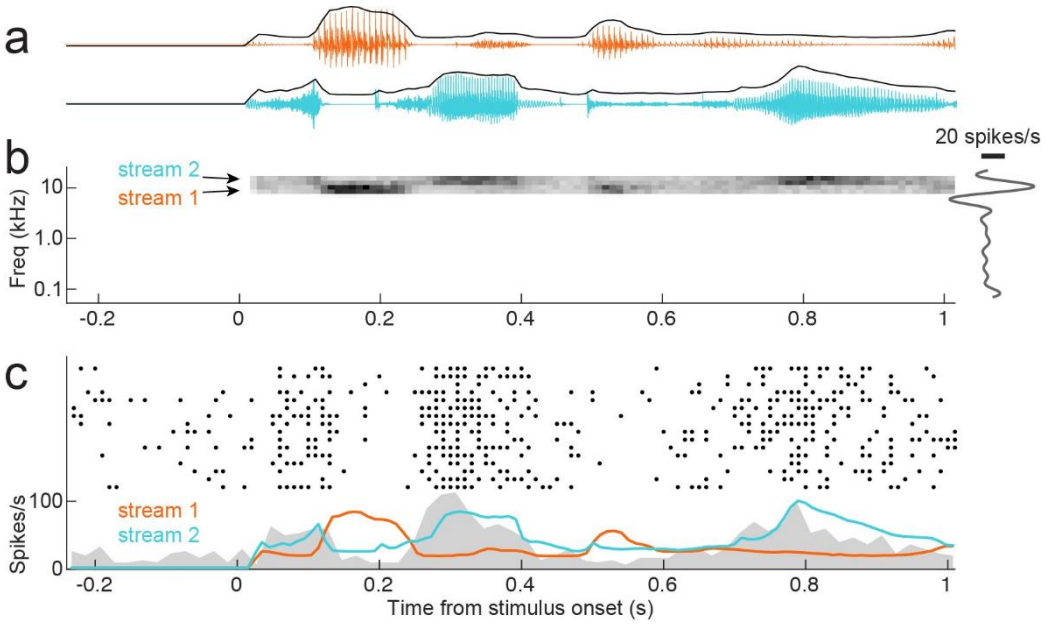

Figure 1
A. Two example natural vocalization waveforms show characteristic interspersed epochs of high sound energy and silence. Each sound has a distinct envelope tracing amplitude over time, which captures these complex temporal dynamics. **B.** Spectrogram of vocalization-modulated noise presented to one A1 neuron. Stimuli were generated by applying vocalization envelopes to narrowband noise, capturing the complex temporal dynamics of natural sounds. For the two-band stimulus, a different envelope was applied to adjacent, non-overlapping spectral bands. Both noise streams were positioned in the responsive area of a frequency tuning curve (right). Thus vocalization-modulated noise enabled probing natural, nonlinear temporal processing while minimizing complexity of spectral features. **C.** Raster response the same neuron to repeated presentations of the vocalization-modulated noise stimulus (top), and peri-stimulus time histogram (PSTH) response averaged across repetitions (gray shading, bottom). The envelope of each noise stream is overlaid. Increased amplitude in stream 2 (blue) leads to a strong onset response that weakens after about 50 ms (transients in the PSTH at 0.25 s and 0.8 s). Stream 1 (orange) suppresses the PSTH, with no evidence for adaptation.

We used a linear-nonlinear spectro-temporal receptive field model (LN model) to establish a baseline characterization of auditory encoding properties (Figure 2 A, (Aertsen & Johannesma, 1981; S. V David et al., 2009; Klein et al., 2000)). This model describes time-varying neural activity as the linear weighted sum of the preceding stimulus spectrogram (Eq. 1). Because the vocalization-modulated noise consisted of just two distinct spectral channels, the model required a filter with only two input spectral channels, compared to multiple spectral channels for analysis of broadband noise or natural sounds. To account for well-established nonlinear threshold and saturation properties of spiking neurons, the linear filter stage was followed by a static, sigmoidal output nonlinearity (Eq. 2, (Thorson et al., 2015), Figure 2 A).

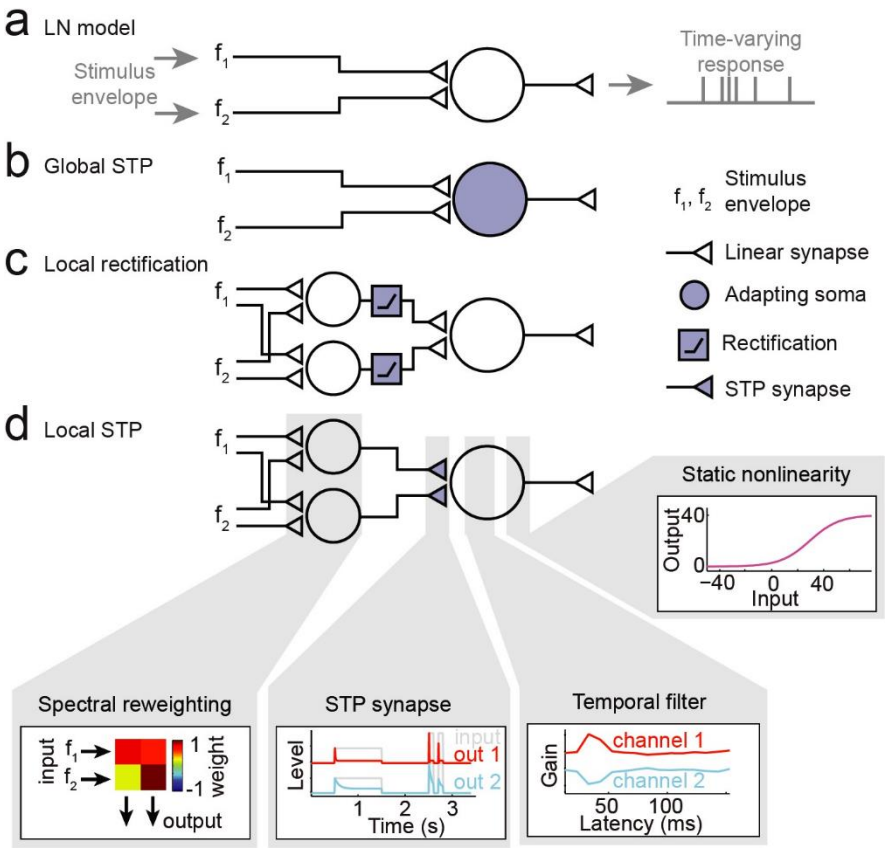

Figure 2 Alternative encoding models to describe auditory neural responses to vocalization-modulated noise.
A. The linear-nonlinear spectro-temporal receptive field (LN model) describes the time-varying neural response as a linear weighted sum of the preceding stimulus envelopes, followed by a static sigmoid nonlinearity to account for spike threshold and saturation. **B.** In the global short-term plasticity (STP) model, nonlinear STP (depression or facilitation) is applied to the output of the linear filter prior to the static nonlinearity. **C.** In the local rectification model, the input channels are linearly reweighted and then nonlinearly thresholded (rectified) prior to the linear temporal filter and static nonlinearity. **D.** In the local STP model, input channels are linearly reweighted, and then nonlinear STP (depression or facilitation) is applied to each reweighted channel, prior to the linear temporal filter and static nonlinearity. Gray boxes show example model parameters applied at each processing stage.

To regularize model fits, we constrained the temporal dynamics of the filter applied to each input channel to have the form of a damped oscillator (Eq. 3, (Thorson et al., 2015)). This parameterization required fewer free parameters than a simple, nonparametric weighting vector and improved performance over the model with a nonparametric linear filter (see Figure 4 C). However, the shape of a single parameterized temporal filter did not capture temporal responses dynamics fully for all neurons. To support more flexible temporal encoding, we introduced a spectral reweighting in which the two input channels were mapped to *J* channels prior to temporal

filtering, with the possibility that $J > 2$ (Eq. 4). Each reweighted input was passed through a separately-fit temporal filter. Several values of J were tested. For the majority of model comparisons, $J = 5$ was found to produce the best performing models, on average, and most results below are for models with this channel count (although $J = 2$ spectral channels achieved nearly asymptotic performance for the LN model, see below).

The LN model, as well as the other models discussed below, was fit using gradient descent (Byrd et al., 1995; Pennington & David, 2022). Model fits were regularized by the parametric formulation of the linear filter (Eqs. 3-4) and by a shrinkage term applied to the mean squared error cost function (Thorson et al., 2015). Model performance was assessed by the accuracy with which it predicted the time-varying response to a novel validation stimulus that was not used for estimation (Wu et al., 2006). Prediction accuracy was quantified by the correlation coefficient (Pearson's R) measured between the predicted and actual PSTH response, corrected to account for sampling limitations in the actual response (Hsu et al., 2004). A value of R=1 indicated a perfect prediction, and R=0 indicated random prediction.

The LN model was able to capture some response dynamics of A1 neurons, but several errors in prediction can be seen in example data (Figure 3, Figure 5). In particular, the LN model failed to account for transient responses following stimulus onset (arrows in Figure 3). A previous study showed that, for stimuli consisting of a single modulated noise band, a model incorporating nonlinear short-term synaptic plasticity (STP) prior to the temporal filtering stage provides a more accurate prediction of neural activity (S. V David & Shamma, 2013). STP is widespread across cortical systems, making it a plausible mechanism to support such adaptation (S. V David & Shamma, 2013; Tsodyks et al., 1998). Given that STP occurs at synaptic inputs, this observation suggests that A1 neurons can undergo adaptation independently for inputs in different spectral channels. Spectrally tuned adaptation could give rise to a rich code for complex spectro-temporal patterns (Buonomano & Maass, 2009). However, based on previous results, it is not clear whether the nonlinear adaptation occurs primarily after information is summed across spectral channels

(global adaptation) or if it occurs separately for the different spectral channels (local adaptation). To determine whether adaptation occurs pre- or post-spectral integration, we estimated two variants of the LN model, a global STP model, in which input spectral channels undergo the same adaptation prior to linear filtering (Figure 2 B), and a local STP model, in which each channel adapts independently according to the history of its own input (Figure 2 D).

Spectral reweighting was applied to the stimulus for STP models (Eq. 4), as in the case of the LN model, above. For the local STP model, nonlinear adaptation occurred after spectral reweighting. The reweighting made it possible for the same band of the vocalization-modulated noise to undergo adaptation at multiple timescales and, conversely, for different bands to be combined into a single channel before adaptation. This flexible arrangement models cortical neurons, where inputs from peripheral channels can be combined either pre- or post-synaptically (X. Gao & Wehr, 2015; Ko et al., 2011). The model schematic shows a model in which the two inputs were reweighted into two channels (Figure 2 D), but we compared models with $J = 1 \ldots 5$ channels (see Figure 4and Methods). As in the case of the LN model, the STP models included the same sigmoidal output nonlinearity. The linear filter and static nonlinearity architectures were the same across LN and STP models, and all models were fit using identical data sets. However, the free parameters for each model were fit separately.

To test for the possibility that any benefit of the local STP model simply reflects the insertion of a nonlinearity into the LN model between spectral reweighting and temporal filtering, we considered an additional model, the local rectification model, in which each reweighted channel was linearly rectified prior to temporal filtering (Eq. 9, Figure 2 C).

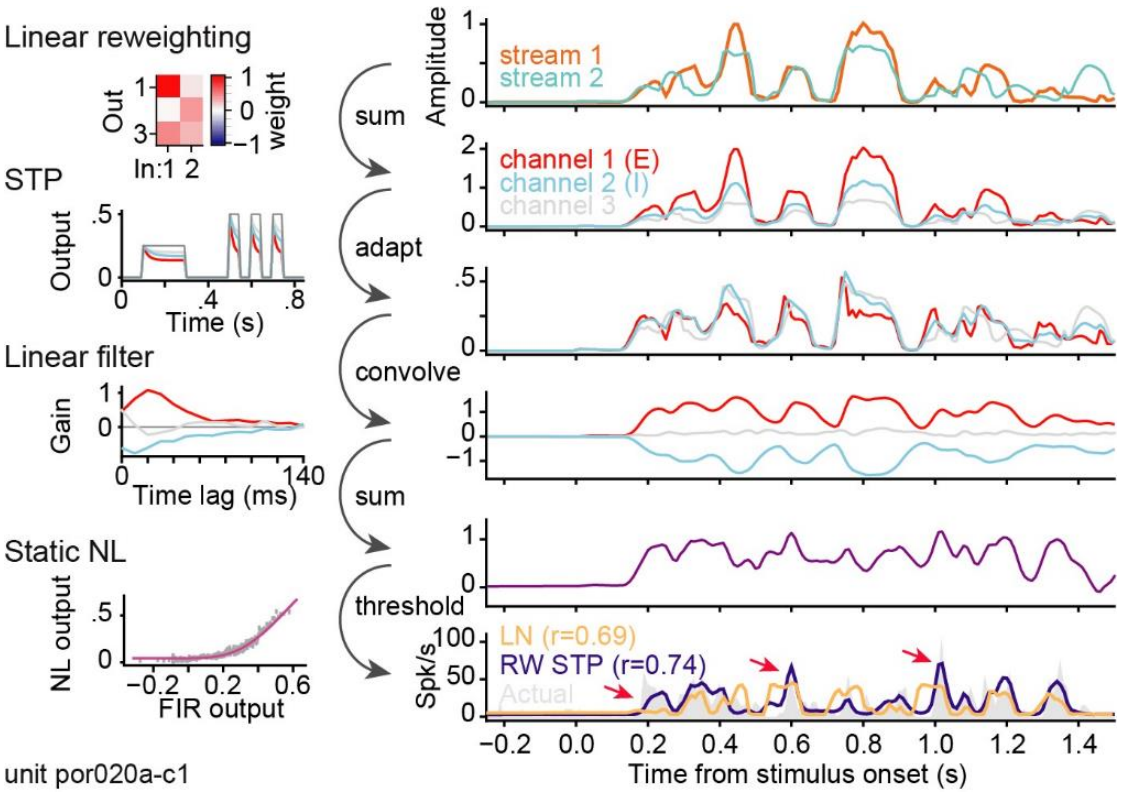

Figure 3
Transformation applied to incoming vocalization-modulated noise for a local short-term plasticity (STP) model estimated for one A1 neuron. Spectral reweighting emphasizes input stream 1 in channel 1 (red), stream 2 in channel 2 (blue), and both streams in channel 3 (gray). All three reweighted channels undergo independent STP. For this neuron, STP is stronger for channel 1 than for the other channels. The linear filter produces excitation for channel 1 and inhibition for channel 2. After the final static nonlinearity (NL), the predicted PSTH (bottom panel, purple) shows a good match to the actual PSTH (gray shading), while the prediction of the LN model does not predict the response dynamics as accurately (orange). Arrows indicate transient PSTH features captured better by the STP model.

These different encoding models can each be cast as a sequence of transformations, where the output of one transformation is the input to the next. Their modularity enables visualization of how the data is transformed at each step of the encoding process. Figure 3 illustrates the transformations that take place in an example local STP model for an A1 neuron (J = 3 spectral reweighting channels shown for simplicity). The vocalization-modulated noise envelope is first linearly reweighted into three channels. In this example, the first reweighted channel closely follows the first input channel. Second, the three reweighted channels undergo independent STP-like adaptation. The first channel experiences the strongest adaptation (red). The adapted channels are then convolved with a linear filter, which in this case is excitatory for channel 1,

inhibitory for channel 2 (blue), and transient excitation for channel 3 (gray). The convolved channels are summed and then passed through a static nonlinearity to generate the final predicted time-varying spike rate. The PSTH response predicted by the reweighted STP model can be compared directly to the actual PSTH and predictions by other models (Figure 4).

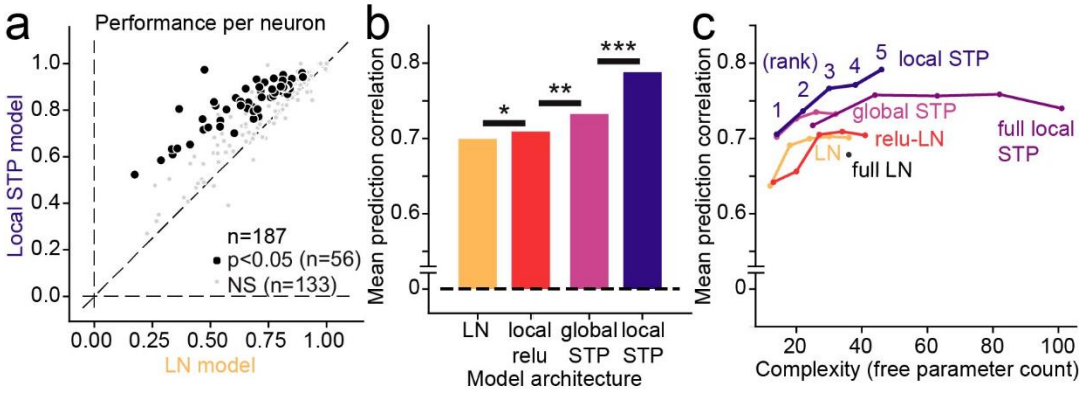

Figure 4
A. Scatter plot compares noise-corrected prediction correlation between the linear-nonlinear (LN) model and local short-term plasticity (STP) model for each A1 neuron. Black points indicate the 56/187 neurons for which the local STP model performed significantly better than the LN model (p<0.05, jackknifed t-test). **B.** Mean performance (noise-corrected correlation coefficient between predicted and actual PSTH) for each model across the set of A1 neurons. The global STP model showed improved performance over the LN and local rectification (relu) model. The local STP model showed a further improvement over the global STP model (*p < 0.01, **p < 10^{-4}, ***p < 10^{-6}, Wilcoxon sign test, n = 187/200 neurons with above-chance prediction correlation for any model). The best performing model, the local STP model, reweighted the two input envelopes into five spectral channels, each of which underwent independent STP prior to linear temporal filtering and a static nonlinearity. **C.** Pareto plot compares model complexity (number of free parameters) versus average prediction correlation for model architectures with and without STP, with and without parameterization of the temporal filter (full vs. DO) and for variable numbers of reweighted spectral channels (rank). Models with STP showed (purple, blue) consistently better performance than models without STP (orange, red) for all levels of complexity.

For 187 out of the 200 A1 neurons studied, at least one model (LN, global STP, local rectification, local STP, J = 5 spectral reweighting channels for all models) was able to predict time-varying responses with greater than chance accuracy (p<0.05, Bonferroni-corrected permutation test). Prediction correlation for the global STP model was significantly greater than the linear model for a subset of neurons (n = 22/187, p < 0.05, permutation test, Figure 4 B). The average noise-corrected prediction correlation across the entire sample of neurons was greater for the global STP model (mean 0.699 vs. 0.732, median 0.715 vs. 0.755, p = 2.1 x 10^{-7}, sign test). Mean performance tended to be slightly lower than median, probably because performance

was near the upper bound of r = 1.0, creating a slight negative bias in the mean. However, we saw no qualitative difference between these metrics in any model comparison. The local rectification model also showed an average improvement in performance over the LN model (mean 0.699 vs. 0.709, median 0.715 vs. 0.732, p = 0.042, sign test, Figure 4 B). However, the local STP model consistently performed better than all the other models (mean 0.795, median 0.818, $p<10^{-8}$ for all models, sign test, Figure 4 B). Prediction accuracy was significantly greater than the LN model for 58/187 neurons ($p < 0.05$, permutation test, Figure 4 A). Taken together, these results indicate that the spectrally tuned nonlinear adaptation described by the local STP model provides a more accurate characterization of A1 encoding than LN models or models in which the adaptation occurs uniformly across spectral channels.

While the local STP model consistently performed better than the other models, its performance could be attributed to its additional complexity, *i.e.*, the fact that it required more free parameters than the other models, rather than something specific about spectrally tuned adaptation. To characterize the interaction of model complexity and performance, we compared prediction accuracy for models with variable numbers of spectral reweighting channels, J=1…5 (Figure 4 C). When compared in a Pareto plot, the local STP model shows a consistent pattern of improved performance over LN models, independent of spectral channel count or overall parameter count. This comparison also included models in which the temporal filter was either parameterized by a damped oscillator (Eq. 3) or nonparameterized ("full"). The parameterized models performed consistently as well or better than their nonparameterized counterparts, indicating that this reduction in dimensionality preserved important temporal filter properties. Thus, the benefit of incorporating local STP is consistent, regardless of model complexity.

Spectrally tuned adaptation is stronger for excitatory than inhibitory inputs

We studied properties of the LN and STP models in order to understand what features of the STP models lead to their improved performance. Response dynamics varied across A1 neurons, sometimes emphasizing only sound onsets and in other cases tracking one or both envelopes

across the entire trial. For many neurons, both models were able to capture the coarse response dynamics, but the STP model was able to predict the transient responses and the relative amplitude of responses more accurately (Figure 5 B-C). In some cases the LN and STP models performed equivalently, indicating that some neurons showed little or no nonlinear adaptation (Figure 5 G).

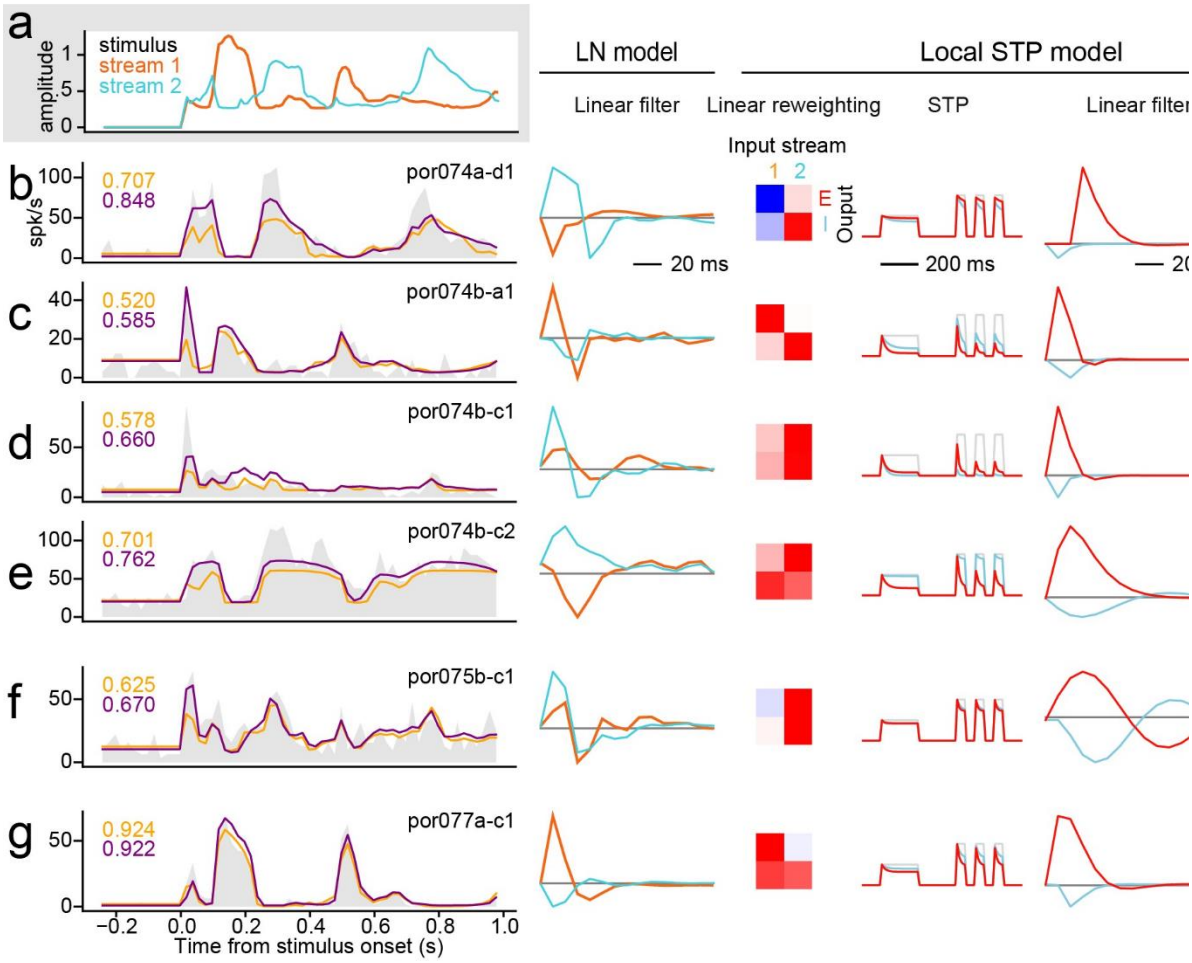

Figure 5
A. Envelope of vocalization-modulated noise streams. **B-G.** Left column, example PSTH responses of several A1 neurons (gray shading). The spectral position of noise bands was adjusted to fall within the receptive field of each neuron, but the envelopes were the same for each recording. Responses were sometimes dominated by one stream (e.g., unit B tracks stream 1 and G tracks stream 2), but could also track both (e.g., unit F). Response dynamics also vary substantially, from sustained, following the stimulus envelope (G), to highly transient responses that attenuate after sound onset (D). Numbers at upper left of PSTH plots indicate prediction correlation for the linear-nonlinear (LN) model (orange) and local short-term plasticity (STP) model (blue). Predicted PSTHs are overlaid on the actual PSTH. Second column shows linear filters from the LN model for each neuron, whose gain reciprocates the PSTH responses. Columns at right show spectral weights, STP properties and linear filters for the largest

(positive gain, red) and smallest (negative gain, blue) temporal filter in local STP models for the same neurons.

Although isolated stimuli in both input channels usually evoked excitatory responses (Figure 1 C), the gain of one filter in both LN and STP models was often negative (Figure 5, middle and right columns). These suppressive responses likely reflect the unmasking of inhibition by broadband stimuli (Eggermont, 2011). The fit procedure was not constrained to require a negative channel, so the presence of negative channels is the result of optimizing model parameters for prediction accuracy. We quantified the gain of each local STP model channel by summing temporal filter coefficients across time lags. By definition, one channel always had the largest gain, which we identified as the strongest input channel. A comparison of gain for largest versus smallest gain showed that one channel was always positive (n = 187/187 units, Figure 6 B). Strikingly, almost every filter contained at least one channel with negative gain (n = 175/187). We focus on models with J=5 spectral reweighting channels here, but nearly the same results are observed for models with J = 2…5. There was no difference in the prevalence of inhibitory channels in neurons that showed a significant improvement for the STP model, compared to neurons that did not show an improvement (p > 0.05, unpaired t-test, n=56/175 improved, n=119/175 not improved, see Figure 6 A and below). Although the specific mechanisms producing positive and negative gain are not determined in this model, we refer to them as excitatory and inhibitory gain, respectively.

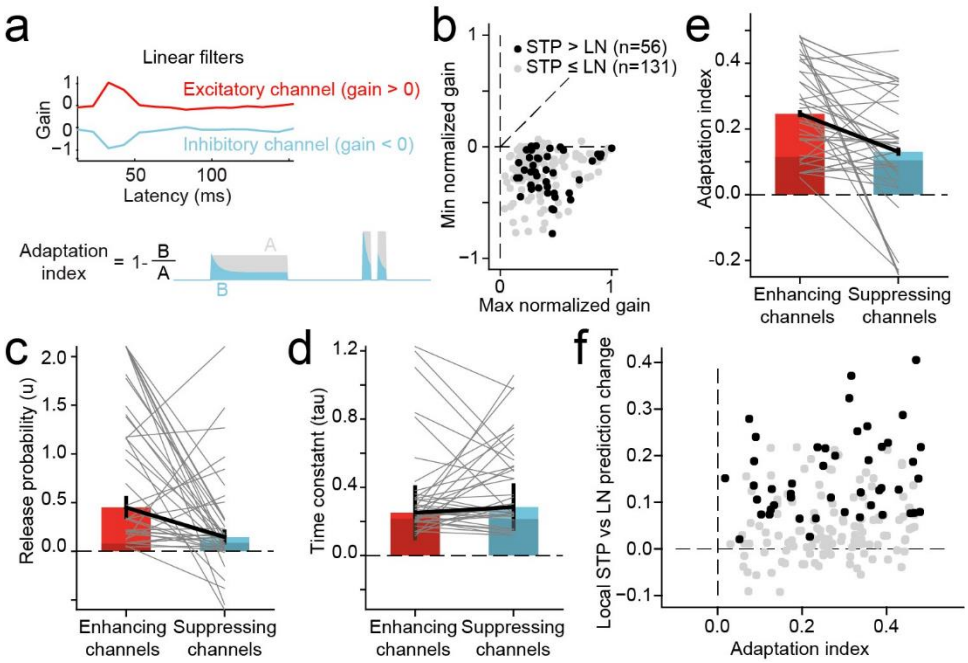

Figure 6
A. In the local STP model, each reweighted spectral channel passed through a nonlinear filter, mimicking synaptic STP, a nonlinear transformation, prior to the linear temporal filter stage. An index of adaptation strength for each model synapse was computed as one minus the fraction change in the amplitude of a test signal after passing through the adapting synapse. An index value > 0 indicated synaptic depression, and a value < 0 indicated facilitation. **B.** Overall gain for each channel of the linear filter in the STP model was computed as the sum of the filter across time lags. Scatter plot compares gain for the channel with largest magnitude, which was always positive (horizontal axis), and for the channel with smallest magnitude, which was either positive or negative (vertical axis). The vast majority of model fits contained at least one excitatory (positive) channel and one inhibitory (negative) channel (n = 183/187). Units in which the local STP model generated a significant improvement in prediction power are colored black (n = 56, $p<0.05$, jackknife *t*-test). **C.** Comparison of release probability parameter fit values for STP filters in excitatory versus inhibitory channels (n = 56 STP models with significant improvement in prediction power). Gray lines connect values for a single model. Average values were significantly greater for excitatory versus inhibitory synapses for release probability (mean 0.45 vs. 0.15, $p = 1.4 \times 10^{-6}$, sign test). **D.** Comparison of STP recovery time constant, plotted as in D, shows no difference between excitatory and inhibitory channels (mean 0.063 vs. 0.081 s, $p > 0.5$, sign test). **E.** Comparison of adaptation index shows a significant difference between excitatory and inhibitory channels (mean 0.25 vs. 0.13, $p = 2.8 \times 10^{-4}$ sign test). **F.** Scatter plot compares average adaptation index for each local STP model against the change in prediction correlation between the LN and local STP model. There is a positive correlation between STP effects and changes in prediction accuracy (r = 0.17, p = 0.023, n = 187, Wald Test). Neurons with significant changes in prediction accuracy are plotted as in B.

We wondered whether adaptation captured by the STP model differed between excitatory and inhibitory channels. For the 56 neurons with improved performance by the local STP model (see Figure 4, above), we compared STP parameters (release probability and recovery time constant, see Eq. 6) and the overall adaptation index between highest- and lowest gain channels. The adaptation index was measured as one minus the ratio of the output to input of the synapse for a

standard test input (Figure 6 A, (S. V David & Shamma, 2013)). Index values greater than zero indicated depression, and values less than zero indicated facilitation. When we compared STP properties between channels, we observed that release probability and adaptation index were both stronger, on average, for excitatory versus inhibitory channels ($p = 0.0011$ and $p = 4.3 \times 10^{-4}$, respectively, sign test, Figure 6 C, E). The mean adaptation index of excitatory channels (0.27) was more than twice that of inhibitory channels (0.13). These results suggest that excitatory responses in A1 tend to adapt following sustained input, while concurrent inhibition undergoes little or no adaptation. Mean recovery time constant did not differ between excitatory and inhibitory channels, possibly because the value of the time constant has little impact on model behavior when adaptation is weak (Figure 6 D).

We also tested whether the magnitude of STP-like adaptation predicted the relative performance of the local STP model. A comparison of average adaptation index versus change in prediction accuracy between the LN and local STP model for each neuron shows a small but significant correlation ($r = 0.17$, $p = 0.023$, $n = 187$, Wald Test for non-zero slope, Figure 6 F). When we considered neurons for which local STP model performance was not greater than the LN model, no mean difference was observed between excitatory and inhibitory channels (Figure 6 C-E, dark bars). However, the local STP models did tend to show non-zero STP strength, even if there was no significant improvement in performance. While many neurons did not show a significant improvement in prediction accuracy for the local STP model, the vast majority showed a trend toward improvement (168/187, Figure 4 A). If more data were available, permitting more robust model estimates, the number of neurons showing significant STP effects could be larger.

Spectrally tuned adaptation supports contextual effects of stimulus-specific adaptation

Nonlinear adaptation has previously been proposed to play a role in contextual effects on auditory cortical responses (Asari & Zador, 2009). One common measure of contextual influences on auditory activity is stimulus specific adaptation (SSA, (Pérez-González & Malmierca, 2014; Ulanovsky et al., 2003)). When two discrete stimuli are presented in a regular sequence, with a

standard stimulus presented more frequently than an oddball stimulus, responses to the standard tend to undergo adaptation, but responses to the oddball stimulus can be less adapted or even facilitated relative to a silent context. Effects of SSA have been attributed to feedforward adaptation and/or lateral inhibition (Carbajal & Malmierca, 2018; Natan et al., 2015; Nelken, 2014; Yarden & Nelken, 2017).

To test for SSA effects, for a subset of neurons we presented standard/oddball sequences of noise bursts, falling in the same spectral bands as the vocalization-modulated noise stimuli. We measured SSA for these responses by an SSA index (SI) that compared responses to noise bursts when they appeared as standards vs. oddballs (Ulanovsky et al., 2003). Adaptation effects were weaker than previously been reported for A1 in anesthetized animals, but SI was significantly greater than zero in 43% of neurons (p<0.05, standard/oddball permutation test, n = 44/102). We tested whether models including STP could predict responses to oddball stimuli and explain SSA effects. LN, global STP, and local STP models were fit to data collected during the presentation of the oddball sequences. Because the design of the oddball stimulus experiments did not include repetitions of the same sequences, models were fit and tested using single trials. This design precluded correcting the prediction correlation for variability in the neural response (Hsu et al., 2004; Thorson et al., 2015), leading to comparatively lower correlation values than for the other stimulus sets. Nonetheless, the introduction of nonlinear model elements improved prediction accuracy. Between the LN model vs. local STP model, 46% of the cells responses were significantly better predicted by the local STP model ($p < 0.05$, jackknifed t-test, n = 47/102, Figure 7 A). Each model showed a significant improvement in accuracy over the simpler one (LN vs global STP model, $p = 1.7 \times 10^{-4}$; global STP model vs local STP model, $p = 1.9 \times 10^{-13}$, sign test, Figure 7 C). This pattern of improvement closely parallels the vocalization-modulated noise data (Figure 4).

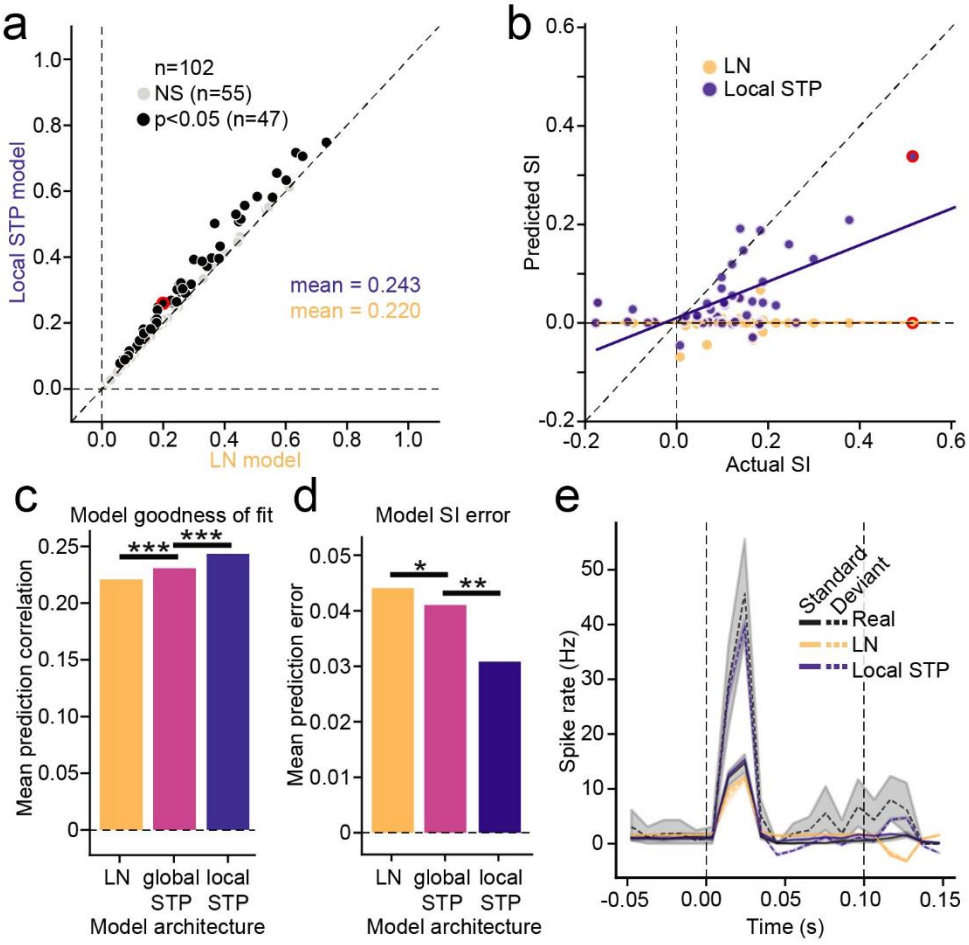

Figure 7

A. Scatter plot compares prediction accuracy for the LN model and local STP model, estimated using oddball stimuli for each neuron. Black markers indicate significant difference in the performance between models (p<0.05, jackknifed t-test). **B.** Scatter plot compares SSA index (SI) calculated from actual responses against SI from responses predicted by LN model (orange) and local STP model (blue) for neurons with significant actual SI (p < 0.05, standard/oddball permutation test). The LN model is unable to account for any stimulus specific adaptation, while the SI predicted by the local STP model is correlated with the actual values (LN: r = 0.011, p = 0.95; local STP: r = 0.636, p = 3.4 x 10^{-6}, Wald Test for non-zero slope). **C.** Summary of the mean prediction correlation for all cells across all tested models (LN model vs. global STP model, p = 1.7 x 10^{-4}, global STP model vs. local STP model, p = 1.9 x 10^{-13} , LN model vs local STP model, p = 1.1 x 10^{-15}, sign test). **D.** Mean SI prediction error for each model architecture. The prediction error for each cell is the mean standard error (MSE) between actual and predicted SI (LN model vs. global STP model, p = 0.024; global STP model vs. local STP model, p = 0.005, LN model vs. local STP model, p = 1.5 x 10^{-4}, sign test). **E.** Example actual (black), LN model-predicted (yellow) and local STP model-predicted (blue) PSTH response to standard (continuous line) and deviant (dashed line) noise bursts. Shaded areas standard error on the mean (bootstrap p = 0.05). Vertical lines mark sound onset and offset. For the LN model, both standard and oddball predictions are close to the actual standard response, but the local STP model predicts the enhanced oddball response. Example cell is highlighted in red in panels A and B. *p <0 .05, **p < 0.01, ***p < 0.001.

Because the models could predict time-varying responses to the noise stimuli, we could measure SI from responses predicted by the models. For neurons with significant SI (p < 0.05,

permutation test, n = 44/102) we measured the correlation between actual and predicted SI values. The best performing model, the local STP model, was able to significantly predict the SI (n = 44, r = 0.636, p = 3.4 x 10^{-6}, Wald Test for non-zero slope, Figure 7 B, blue). On the other hand, the LN model was unable to predict SI (n = 44, r = 0.011, p = 0.95, Wald Test, Figure 7 B, orange). When comparing the SI prediction error across model architectures, the mean population error consistently decreased with the addition of spectrally tuned adaptation (LN vs global STP model, p = 0.024; global STP model vs local STP model, p = 0.005, sign test, Figure 7 D). Thus, A1 neurons that showed evidence for nonlinear STP-like adaptation also exhibited SSA, indicating that the two phenomena may share common mechanisms.

Nonlinear adaptation is robust to changes in behavioral state

Several previous studies have shown that the response properties of neurons in A1 can be affected by changes in behavioral state. When animals engage in a task that require discrimination between sound categories, neurons can shift their gain and selectivity to enhance discriminability between the task-relevant categories (Fritz et al., 2003; Niwa et al., 2012; Otazu et al., 2009). Changes in overall gain are observed most commonly. Effects on sensory selectivity have been more variable and difficult to characterize.

We tested if changes in behavioral state influence the nonlinear STP-like adaptation we observed in A1. We trained ferrets to perform a tone detection task, in which they reported the occurrence of a pure tone target embedded in a vocalization-modulated noise sequence (Figure 8 A). We recorded neural activity during passive listening to the task stimuli and during active performance of the tone detection task. We then estimated STP models in which the model parameters were either fixed between behavior conditions (passive listening versus active behavior) or allowed to vary between conditions. Because identical stimuli were used in both conditions, differences in the model fit could be attributed to changes in behavioral state. As in the case of SSA data, noise stimuli were not repeated within a behavioral block. Thus prediction

accuracy was assessed with single trial data, and absolute prediction measures were lower than for the passive data reported above (Figure 4).

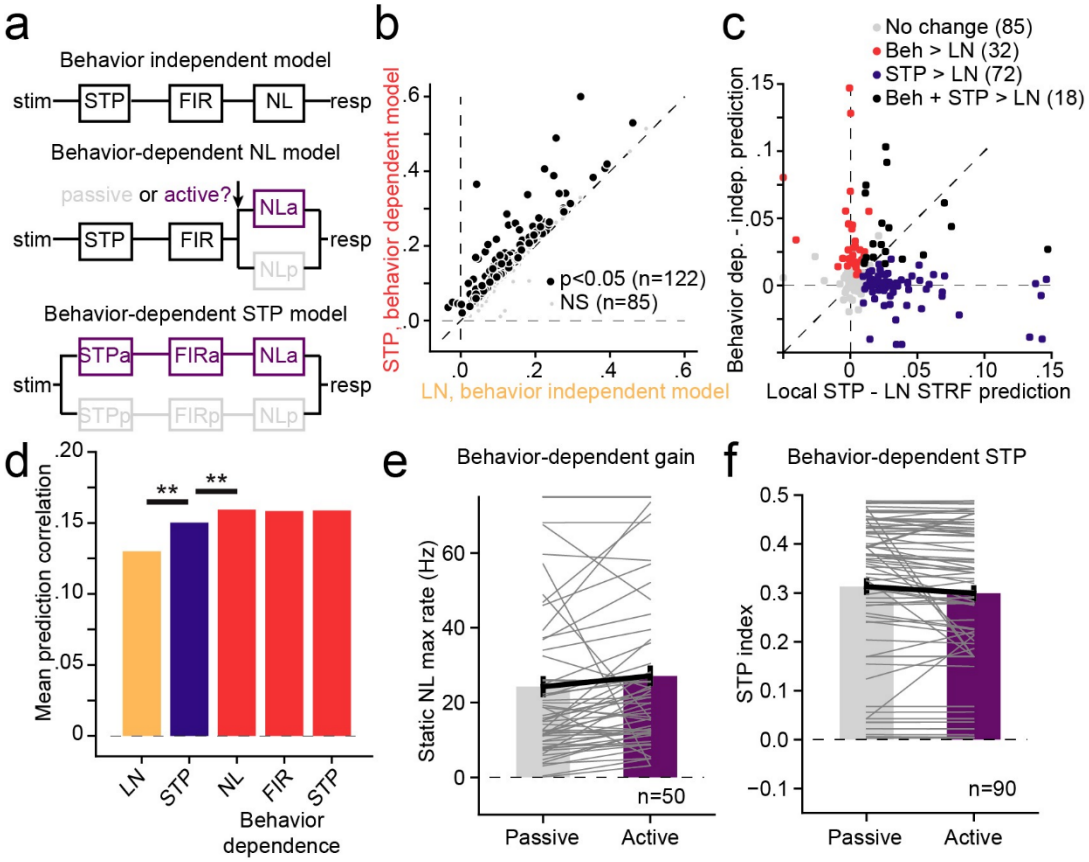

Figure 8

A. Schematic of alternative behavior-dependent local STP models that account for changes in sound encoding between passive and active tone detection conditions. The behavior-independent model was fit independent of behavior state. For the behavior-dependent NL model, the static nonlinearity was fit separately for passive and active conditions but all other parameters were constant. Subsequent models introduced the active/passive split prior to earlier stages. **B.** Scatter plot compares prediction accuracy between the behavior-independent LN model and full behavior-dependent STP model for each cell in the set (pooled across on BF and away from BF target blocks). 122/207 neurons show a significant increase in prediction accuracy for the behavior-dependent model ($p<0.05$, jackknifed t-test). **C.** Relative change in prediction accuracy for each neuron from incorporating STP (LN vs. local STP model, x axis) versus incorporating behavior dependence (behavior independent vs. –dependent, y axis). The small number of units that show improvement for both models (black), is in the range expected by chance if STP and behavior effects are distributed independently across the A1 population ($p>0.2$, permutation test). **D.** Comparison of mean prediction accuracy for each model reveals a significant increase in performance for STP model over the LN model, as in the passive-only dataset in Figure 4 (mean 0.13 vs. 0.15, $p < 10^{-10}$). In addition, for the STP model, the behavior-dependent NL model shows improved performance over the behavior-independent model (mean 0.150 vs. 0.159, $p = 2.2 \times 10^{-7}$, sign test). However, no further improvement is observed if the linear filter or STP parameters are made behavior-dependent ($p>0.05$, sign test). **E.** Comparison of passive vs. active model gain (amplitude of the static nonlinearity) between active and passive conditions shows an increase in the mean response during behavior (mean NL amplitude 24 vs. 27 spikes/sec, $p = 2.0 \times 10^{-5}$, sign test). Gray lines show passive vs. active amplitude for each neuron. **F.** Comparison STP index for behavior-dependent model shows a small decrease in

STP in the activity condition (mean 0.31 vs. 0.30, p = 0.002). This small behavior-dependent change does not impact mean prediction accuracy, as plotted in D.

When the parameters of the static nonlinearity were allowed to vary between passive and active states, allowing changes in gain between the passive and activate conditions, the models showed a significant improvement in predictive power when compared to the behavior-independent model (mean single trial prediction correlation 0.13 vs. 0.15, $p = 7.1 \times 10^{-13}$, sign test, Figure 8 B-D). However, allowing other model parameters to vary with behavioral state provided no additional improvement in model performance ($p > 0.05$, sign test, Figure 8 D). Thus, the changes in behavioral state influence the overall gain of the neural response without affecting the linear filter or nonlinear adaptation captured by the STP model.

We also considered whether the presence of STP-like adaptation in a neuron predicted its tendency to show behavior-dependent changes in activity. When we compared the incremental change in prediction accuracy resulting from addition of nonlinear STP or behavior-dependent gain to the encoding model, the relationship was highly variable (Figure 8 C). Some neurons showed improvement only for STP or behavior-dependence, and just a small number showed improvements for both. Overall, these effects occurred independently across the population ($p >$ 0.1, permutation test). Thus the improved performance of the STP model does not predict the occurrence of behavior-dependent changes in activity.

The comparison of prediction accuracy between behavior-dependent models suggests that the response gain can change between passive and active conditions but STP parameters do not. When we compared parameters between models fit separately under the different behavioral conditions, we found this to be largely the case. The average gain of the auditory response increased when animals engaged in behavior (mean amplitude of static nonlinearity: 24 vs. 27 spk/sec, $p = 2 \times 10^{-5}$, n = 50 neurons with significant improvement in behavior-dependent vs. behavior-independent model, sign test, Fig 8E). The average STP index showed a small decrease during task engagement (mean STP index 0.31 vs. 0.30, p = 0.0016, n = 10 neurons with

significant improvement in local STP vs. LN model, sign test, Figure 8 F, right panels). While this change in STP index was significant, allowing it to fluctuate did not significantly impact prediction accuracy (Figure 8 D). A larger dataset may uncover significant influences of behavior-dependent nonlinear adaptation. However, the current analysis suggests that changes in STP play an overall smaller role in mediating behavioral effects in A1 than changes in overall response gain (Figure 8 E).

Natural stimuli reveal nonlinear adaptation of spectrally overlapping channels in A1

The vocalization-modulated noise data reveal that spectrally distinct inputs can undergo independent adaptation in A1, supporting contextual coding phenomena such as SSA. In order to understand these nonlinear adaptation effects in a more ethological context, we also recorded the activity of 499 A1 neurons from 5 awake, passive ferrets during presentation of fully natural sounds. The natural stimuli were drawn from a large library of natural sounds (textures, ferret vocalizations, human speech and recordings of the ambient laboratory environment), chosen to sample a diverse range of spectro-temporal modulation space.

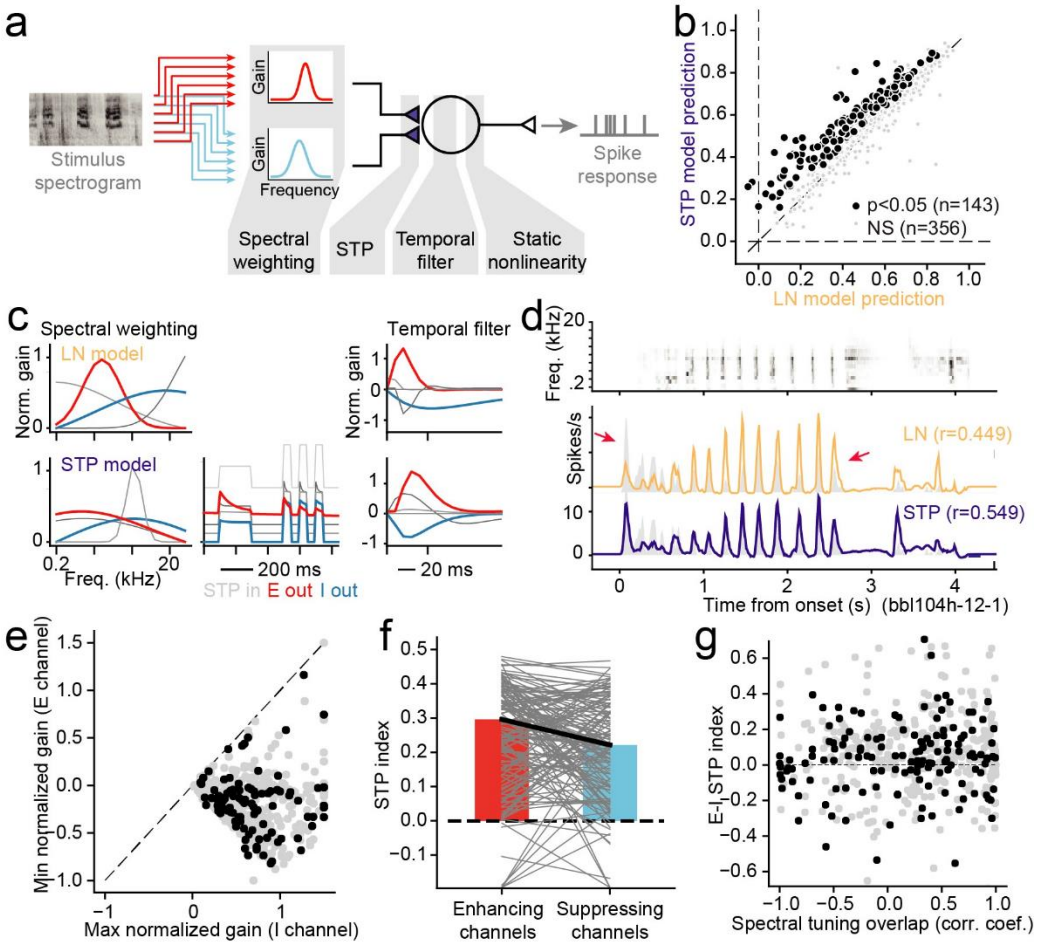

Figure 9 Performance of a local STP model on for A1 encoding of natural sounds.
A. The encoding model for natural sounds resembled reweighted STP model for vocalization-modulated noise, except that the spectral filters at the first stage were two independently fit Gaussian functions that required two free parameters each (mean, standard deviation) and provided a simple tuning function for each spectral channel. **B.** Scatter plot compares prediction accuracy between the LN model and local STP model for the natural sound data. Across the entire set, 143/499 neurons showed a significant improvement in prediction accuracy for the local STP model ($p < 0.05$, jackknife *t*-test). Mean prediction accuracy for the local STP model was significantly greater than the LN model (0.517 vs. 0.563, $p < 10^{-10}$, sign test). **C.** Example spectral weights and temporal filters for one LN model (top) and spectral weights, STP, and temporal filters for the local STP model for the same neuron. Maximum gain is normalized to 1, but the relative gain between channels is preserved. As is typical in the vocalization-modulated noise data, the highest gain filter (red) shows relatively strong STP, and the lowest gain (blue) shows weaker STP. **D.** Predicted PSTH responses for each model for one natural sound stimulus, overlaid on the actual PSTH (gray). The LN model prediction (orange) undershoots the initial transient response and over-predicts the sequence of transient responses later in the stimulus (arrows), while the STP model predicts these features more accurately (blue). **E.** Comparison of gain for the most positive (max normalized gain) and most negative (min normalized gain) linear filter channels for STP models reveals that the majority fits contain one excitatory and one inhibitory channel. **F.** Comparison of STP strength between excitatory and inhibitory channels shows consistently stronger depression for the excitatory channels (mean 0.30 vs. 0.22, $p = 4.1 \times 10^{-3}$, sign test, $n = 143$ units with significant improvement for the STP model). **G.** Scatter plot compares overlap of E and I spectral channels for each STP model (x axis) and relative difference in STP index between the E and I channels. There is no

correlation between tuning overlap and STP index difference, suggesting that A1 neurons represent incoming sound with a diverse combination of spectral tuning and nonlinear adaptation.

Neural encoding properties were modeled by a reduced-rank model (Figure 9 A, (Simon et al., 2007; Thorson et al., 2015)). For the LN model, the sound spectrogram passed through a bank of J = 4 spectral filters, each of which computed a linear weighted sum of the spectrogram at each time bin. The spectral filter output then passed through a linear temporal filter (constrained to be a damped oscillator, Eq. 3) and static nonlinearity, identical to elements in the vocalization-modulated noise models. To test for nonlinear adaptation, local STP was introduced to the model following the spectral filtering stage (Figure 9 A).

The STP model predicted time-varying natural sound responses more accurately, on average, than the LN model (Figure 9 B). The STP model performed significantly better for 143/499 of the A1 neurons studied, and the average prediction accuracy was significantly higher for the STP model (mean noise-corrected prediction correlation 0.517 vs. 0.563, median: 0.540 vs. 0.583, $p < 10^{-20}$, sign test). Thus, introducing local nonlinear adaptation to a spectro-temporal model for encoding of natural sounds provides a similar benefit as for encoding of vocalization-modulated noise.

An example comparing LN and local STP model fits for one neuron shows a similar pattern of spectrally tuned adaptation as observed for the vocalization-modulated noise data (Figure 9 C). In this example, the spectral channel with strongest positive gain (red) shows relatively strong STP, while the channel with strongest negative gain (blue) shows very little evidence for STP. The net effects of this tuned STP can be observed in the predicted PSTH response to a natural sound (Figure 9 D). The LN model fails to predict the strong transient response at the sound onset and over-predicts the sequence of transients 1-2 sec after sound onset. The local STP model captures these dynamics more accurately.

As in the case of the vocalization-modulated noise data, we compared STP effects between excitatory and inhibitory channels. Temporal filters were ordered by their average gain, and the

highest- and lowest gain filters were selected for comparison of STP properties (Figure 9 E). This comparison revealed that mean STP index was significantly larger for excitatory channels (mean 0.30) than for inhibitory channels (mean 0.22, $p = 4.1 \times 10^{-3}$, sign test, Figure 9 F). As in the case of vocalization-modulated noise (Figure 6), the weaker STP for inhibitory channels suggests that these inputs tend to undergo little or no adaptation, while excitatory inputs undergo stronger adaptation. These effects did not depend on spectral tuning of the filters, as the differences in STP for excitatory versus inhibitory channels were consistent across filter center frequencies. There was also substantial heterogeneity in the strength of STP and the degree of overlap between spectral filters in a model fit (Figure 9 G). Thus, while many A1 neurons showed evidence for STP-like adaptation, especially in excitatory channels, the amount of adaptation and spectral overlap varied widely between neurons.

Discussion

We found that the adaptation of neurons in primary auditory cortex (A1) to natural and naturalistic sounds is spectrally selective. These adaptation effects can be modeled by a neuron with multiple input synapses that independently undergo short-term plasticity (STP). Spectro-temporal receptive field models that incorporate nonlinear, spectrally tuned adaptation predict neural responses more accurately than the classic linear-nonlinear (LN) model for both naturalistic vocalization-modulated noise and for fully natural stimuli. They also predict responses more accurately than models that undergo a global (non-spectrally tuned) adaptation. These adaptation effects are stable across changes in behavioral state, even as neurons undergo task-related changes in the gain of sound-evoked responses (Otazu et al., 2009; Schwartz & David, 2018). While the observed adaptation could be produced by a mechanism other than STP, these results demonstrate a general principle, that spectrally tuned adaptation plays an important role in encoding of complex sound features in auditory cortex. Across a variety of stimulus conditions

(Figure 4, Figure 7, Figure 9) and models of varying complexity (Figure 4, Figure 10), a simple STP-like mechanism provides a consistent improvement in the performance of encoding models.

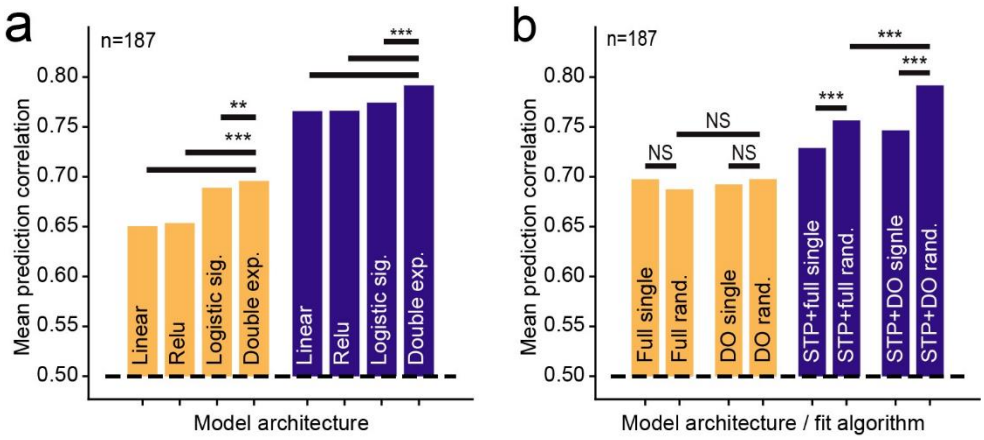

Figure 10
A. Impact of model output nonlinearity on prediction accuracy. Groups of bars compares mean prediction accuracy of LN (orange) and local STP models (blue) with different output nonlinearities using the vocalization-modulated noise data. In both the LN and STP architectures, the double exponential sigmoid shows better performance than a model with no output nonlinearity (linear), linear rectification (relu), and a logistic sigmoid (***$p<10^{-5}$; NS: $p>0.05$ sign test). **B.** Comparison of initialization method and parameterization on LN (orange) and local STP model (blue) performance. Full models used non-parameterized temporal filter functions, and DO indicates model in which the temporal filter is constrained to be a damped oscillator. Single fits started from a single initial condition, and random fits stated a 10 different initial conditions, selecting the best-performing model on the estimation data. Initialization and parameterization had little impact on LN model performance, but both random initialization and DO parameterization improved performance for the local STP model (**$p<10^{-4}$; ***$p<10^{-5}$; NS: $p>0.05$ sign test).

Spectrally tuned adaptation may support perception of complex sound features, such as phonemes in speech and vocalizations (Mesgarani, Cheung, et al., 2014), and may be of particular importance for hearing in noisy environments (Mesgarani, David, et al., 2014; R. C. Moore et al., 2013; Rabinowitz et al., 2013). Evoked activity in A1 rarely has latency longer than 50 ms, but adaptation lasting several tens to hundreds of milliseconds can modulate these short-latency responses. Neurons that undergo adaptation will change their effective spectro-temporal tuning while non-adapting neurons will not. By comparing responses of adapting and non-adapting neurons, a decoder can infer information about stimuli at longer latencies (S. V David & Shamma, 2013). Thus adaptation can operate as an encoding buffer, integrating stimulus information over a longer time window than the latency of the evoked response.

Neural coding of auditory context

Studies of contextual effects on auditory neural coding have shown that the spectro-temporal selectivity can change with statistical properties of sound stimuli, including temporal regularity (Pérez-González & Malmierca, 2014; Ulanovsky et al., 2003), contrast (Dean et al., 2005; Rabinowitz et al., 2012), intensity (Dean et al., 2008; Lesica & Grothe, 2008b; Nagel & Doupe, 2008), and noisy backgrounds (Mesgarani, David, et al., 2014; Rabinowitz et al., 2013). These contextual effects are typically measured in the steady state: neural activity is characterized during discrete epochs in which the statistical properties defining context are held constant. The current results suggest that the same mechanisms that affect activity in the steady state also operate dynamically during the encoding of complex natural stimuli. Spectrally tuned adaptation supports a rich spectro-temporal code in which a continuously changing sensory context, reflecting the previous 100-1000 ms, modulates short-latency (0-100 ms) responses to continuous natural sounds (Ulanovsky et al., 2004).

The timecourse of STP-like adaptation occurs over tens to hundreds of milliseconds, consistent with the timecourse of adaptation in encoding models that incorporate contrast gain control (Rabinowitz et al., 2012). These effects may also share dynamics with models in which local sensory context of synthetic tone stimuli modulates sound-evoked activity (Williamson et al., 2016; Yaron et al., 2012). Previous studies of gain control and contextual modulation have suggested, variously, that either feed-forward adaptation of inputs (in cortex or midbrain), spike-frequency adaptation locally, or recurrent cortical circuits could shape the encoding of spectro-temporal sound features (Keine et al., 2016; Rabinowitz et al., 2012; Ulanovsky et al., 2004; Williamson et al., 2016; Willmore et al., 2016). In all cases, relatively slow changes in stimulus contrast or power around the neurons receptive field can influence sensory selectivity. Thus mechanisms other than STP may be able to support adaptation with similar dynamics. Further study is required to determine if these different models are functionally equivalent or how feedforward and feedback elements of the auditory network contribute to this dynamic coding.

An adaptation-based contextual code produced by mechanisms such as STP may extend broadly across the brain (Buonomano & Maass, 2009; Rothman et al., 2009). As a general computation, this nonlinear adaptation may serve to remove temporal correlations from upstream inputs. Theoretical studies of the visual cortex have argued that variation in synaptic depression across neurons can explain differences temporal frequency tuning across neurons (Chance et al., 1998; Fortune & Rose, 2001). Synaptic depression has also been implicated in producing gain control in hippocampus (Rothman et al., 2009). Thus an auditory code that uses spectrally tuned adaptation provides an example of a computational process that may occur generally across neural systems.

Dynamic reweighting of excitatory and inhibitory input

While the STP model used in this study supported both depression and facilitation, the vast majority of measured adaptation effects were consistent with depression. Moreover, the strength of depression was generally much stronger for spectral inputs that produced an increase rather than decrease in neural firing rate. While the underlying mechanisms producing increases versus decreases in spike rate cannot be fully determined from extracellular recordings, we interpret these components of the model algorithmically as excitatory versus inhibitory responses, respectively. The predominance of adaptation in excitatory channels is consistent with a coding system in which responses to the onset of sound are broadly tuned, but as excitation adapts, the sustained inhibition sculpts responses so that sustained activity is tuned to a narrower set of sound features (Kudela et al., 2018). It has been established that the precise timing and relative strength of inhibition versus excitation can substantially impact tuning in A1 (Wehr & Zador, 2003); thus, dynamic changes in their relative strength during natural sound processing could substantially change encoding properties compared to what is measured in more traditional stimulus paradigms.

The relative tuning, strength, and adaptation properties of excitatory versus inhibitory inputs are not stereotyped, but instead they vary substantially across A1 neurons. In most neurons, the

STP model revealed at least partially overlapping excitatory and inhibitory inputs (Figure 9), consistent with previous work (Froemke et al., 2007; Wehr & Zador, 2003). However, across individual neurons, the best frequency and bandwidth of excitatory channels can be greater or smaller than those of the inhibitory channels. Thus, instead of reflecting a fixed pattern of selectivity, neurons display a diversity of tuning properties that supports a rich code of distinct spectro-temporal patterns. This diversity of synaptic properties may explain the differences in selectivity across A1, including spectro-temporal tuning (Chi et al., 2005), monotonic versus non-monotonic level tuning (Schinkel-Bielefeld et al., 2012; Watkins & Barbour, 2011) and temporal versus rate coding of temporal modulations (Bendor, 2015; L. Gao et al., 2016).

The present study was performed on serial recordings of isolated single units, ignoring possible interactions between neurons that could influence sound coding (See et al., 2018). Simultaneous recordings of neural populations will illuminate the role of adaptation on network connectivity and population dynamics that likely contribute to context-dependent encoding (Pillow et al., 2008; Stringer, Pachitariu, Steinmetz, Carandini, et al., 2019).

Minimal complexity for auditory encoding models

A broad goal motivating this study is to identify the essential computational elements that support nonlinear sound encoding in auditory cortex, in particular, under natural stimulus conditions. While several complex, nonlinear models have been shown to predict auditory neural activity better than the LN model (Ahrens et al., 2008; Atencio et al., 2008; Harper et al., 2016; Kozlov & Gentner, 2016; Rabinowitz et al., 2012; Rahman et al., 2019; Williamson et al., 2016; Willmore et al., 2016), no single model has been adopted widely as a new standard. One reason a replacement has not been identified may simply be that the auditory system is complex and that current data are not exhaustive enough to determine a single model that generalizes across stimulus conditions, species, and behavioral states. Indeed, only a few encoding models have been tested with natural stimuli (Kozlov & Gentner, 2016), and these tests have often been

performed in anesthetized animals (Harper et al., 2016; Rahman et al., 2019; Willmore et al., 2016). In addition to data limitations, proposed models are built around different nonlinear elements, but it is likely that they exist in overlapping functional domains. That is, two different models may both perform better than the LN model because they capture the same adaptation process or nonlinear scaling of response gain. A comprehensive comparison of models using the same natural sound data set will help determine the best performing models and their degree of equivalence. To support such an effort, data from this study is publicly available, and the open source toolbox used for model fitting has a modular design, allowing testing of other model architectures in the same computational framework (Pennington & David, 2022).

The current study took steps for testing an encoding model that have not typically been followed in previous studies. First, the local STP model was tested using multiple different types of stimuli (vocalization-modulated noise, oddball sequences, natural sounds), and it was shown to perform better than the LN and global STP models across stimulus conditions. Second, it compared models of varying complexity. For both low- and high-parameter count models, the addition of a relatively simple STP component provides an improvement in performance. Previous studies have suggested that nonlinear adaptation can improve encoding model performance (Rahman et al., 2019; Willmore et al., 2016). These alternative models are sometimes much higher dimensional than standard LN formulations, and it is not clear how complex a model is required to account for adaptation properties. The current study supports the adaptation hypothesis, but it also shows that the adaptation can be implemented with just a small number of additional free parameters, as long as adaptation occurs independently for input spectral channels.

Stimulus specific adaptation

Stimulus specific adaptation is one of the best-studied contextual effects in auditory cortex (Carbajal & Malmierca, 2018; Pérez-González & Malmierca, 2014; Ulanovsky et al., 2003; Yarden & Nelken, 2017). The STP model developed in the current study is able to account for SSA during

steady state sound presentation. At the same time, the STP model reveals that the same adaptation mechanisms support a broader dynamic code, in which the degree of adaptation is continuously updated to reflect the history of the changing stimulus. This adaptation represents a generalization of SSA, as it does not depend strictly on the regularity the sensory input (Rui et al., 2018). In this way, STP parameters provide a complementary metric to SSA, able to explain nonlinear adaptation for a broader set of stimuli and readily scalable to analysis at a population level.

While nonlinear adaptation and SSA effects are correlated, the strength of this relationship varies across individual neurons. This variability supports the possibility that mechanisms other than STP contribute to SSA. The idea that synaptic depression alone can support SSA has been also disputed because oddball stimuli can sometimes evoked responses that are enhanced relative to those stimuli presented in isolation (Nelken, 2014). However, for a neuron with inhibitory inputs that undergo adaptation, a recurrent disinhibition mechanism could produce enhanced oddball responses (Natan et al., 2015). The data in the current study suggest that inhibitory inputs generally show weaker adaptation than their excitatory partners, which is consistent with other modeling studies (Kudela et al., 2018). However, even inhibitory inputs do tend to undergo some depression, leaving open the possibility that they could explain the enhanced oddball responses during SSA. Inhibitory interneurons in auditory cortex have been shown to contribute to SSA (Natan et al., 2015), but their role in natural sound coding has yet to be characterized.

Robustness of adaptation effects across changes in behavioral state

Studies in behaving animals have shown that gain and selectivity of A1 neurons can be influenced by changes in behavioral state, such as arousal, task engagement, and selective attention (Fritz et al., 2003; Kuchibhotla et al., 2016; Niwa et al., 2012; Schwartz & David, 2018). We observed changes in response gain during task engagement, consistent with this previous work, and incorporating behavior state-dependent gain into the LN model improved prediction

accuracy. However, average adaptation properties did not change across behavioral conditions. Moreover, allowing nonlinear adaptation to vary between behavior conditions did not improve model performance. Thus, STP-like adaptation properties appear to be largely stable across top-down changes in behavioral state. It remains to be seen if they change over longer time scales, but the relative stability of tuning suggests that nonlinear adaptation contributes to a veridical code of sound features in A1 that is selectively gated into high-order, behavior-dependent features in downstream auditory fields (Elgueda et al., 2019).

The approach of incorporating behavioral state variables into sensory encoding models may be useful for integrating bottom-up and top-down coding more broadly (S. V. David, 2018). As sound features take on different behavioral meanings, such as when selective attention is engaged, coding in the auditory system must also shift to represent the behaviorally relevant sound features (Fritz et al., 2007; Mesgarani et al., 2010). A complete understanding of state-dependent changes in sound encoding thus requires models of how neurons change their coding properties in different behavioral states.

Conclusion

How the brain represents complex natural stimuli remains an open question in research across sensory systems. The current study provides evidence that nonlinear adaptation, modeling short-term synaptic plasticity and lasting tens to hundreds of milliseconds, supports a rich code for spectro-temporal sound features in auditory cortex. A simple extension of the classic LN model that allows spectral inputs to undergo independent adaptation provides a consistent improvement in encoding model performance for A1 neurons across a wide range of synthetic and natural stimuli. In addition to providing a more accurate, generalizable encoding model, these findings also provide a framework for linking encoding model analysis to studies of how context influences sound coding.

Methods

Ethics Statement

All procedures were approved by the Oregon Health and Science University Institutional Animal Care and Use Committee (protocol #IP00001561) and conform to standards of the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC).

Animal preparation

Eleven young adult male and female ferrets were obtained from an animal supplier (Marshall Farms, New York). A sterile surgery was performed under isoflurane anesthesia to mount a post for subsequent head fixation and to expose a small portion of the skull for access to auditory cortex. The head post was surrounded by dental acrylic or Charisma composite, which bonded to the skull and to a set of stainless steel screws embedded in the skull. Following surgery, animals were treated with prophylactic antibiotics and analgesics under the supervision of University veterinary staff. The wound was cleaned and bandaged during a recovery period. Starting after recovery from implant surgery (about two weeks), each ferret was gradually acclimated to head fixation using a custom stereotaxic apparatus in a plexiglass tube. Habituation sessions initially lasted for 5 minutes and increased by increments of 5-10 minutes until the ferret lay comfortably for at least one hour.

Acoustic stimulation

Five awake, passively listening, head-fixed animals were presented with vocalization-modulated noise (S. V David & Shamma, 2013; Lesica & Grothe, 2008a; Schwartz & David, 2018) (Figure 1). The stimuli consisted of two streams of narrowband noise (0.25-0.5 octave, 65 dB peak SPL, 3 s duration). Each stream was centered at a different frequency and modulated by a different envelope taken from one of 30 human speech recordings (Garofolo, 1988) or ferret vocalizations from a library of kit distress calls and adult play and aggression calls (S. V David & Shamma, 2013). Envelopes were calculated by rectifying the raw sound waveform, smoothing

and downsampling to 300 Hz. Each envelope fluctuated between 0 and 65 dB SPL, and its temporal modulation power spectrum emphasized low frequency modulations, with 30 dB attenuation at 10 Hz, typical of mammalian vocalizations (Singh & Theunissen, 2003). Thus, the spectral properties of the noise streams were simple and sparse, while the temporal properties matched those of ethological natural sounds.

For three animals, vocalization-modulated noise was presented during passive listening and during active performance of a tone detection task (see below). During passive experiments, both noise streams were positioned in non-overlapping frequency bands in a neuron's receptive field (0.25-1 octave center frequency separation) and were presented from a single spatial location, 30 deg contralateral from the recorded hemisphere. During behavioral experiments, the streams were centered at different frequencies (0.9-4.3 octave separation) and presented from different spatial locations (±30 degrees azimuth), such that one stream fell outside of the spectral tuning curve. Spectral properties of the individual vocalization-modulated noise streams were otherwise identical to those used in the passive experiments above.

In a subset of experiments (two animals), an oddball stimulus was presented to passively listening animals to characterize stimulus-specific adaptation (Ulanovsky et al., 2003). Stimuli consisted of a sequence of regularly repeating noise bursts (100 ms duration, 30 Hz), with the same center frequency and bandwidth as the vocalization-modulated noise presented during the same experiment. On each 20-second trial, 90% of the noise bursts fell in one band (standard) and a random 10% were in the other band (oddball). The spectral bands of the standard and oddball streams were reversed randomly between trials.

Finally, in a different set of experiments, six passively listening animals were presented a library of 93, 3-sec natural sounds. The natural sounds included human speech, ferret and other species' vocalizations, natural environmental sounds, and sounds from the animals' laboratory environment.

In all experiments, the majority of stimuli (28 vocalization-modulated noise samples and 90 natural sounds) were presented a few times (2-5 repetitions). The remaining samples from each sound library (2 vocalization-modulated noise samples and 3 natural sounds) were presented 10-30 times, allowing for robust measurement of a peri-stimulus time histogram (PSTH) response (Figure 2). These high-repeat stimuli were used for measuring model prediction accuracy (see below).

Experiments took place in a sound-attenuating chamber (Gretch-Ken) with a custom double-wall insert. Stimulus presentation and behavior were controlled by custom software (Matlab). Digital acoustic signals were transformed to analog (National Instruments), amplified (Crown D-75A), and delivered through free-field speakers (Manger W05, 50-35,000 Hz flat gain) positioned ±30 degrees azimuth and 80 cm distant from the animal. Sound level was calibrated against a standard reference (Brüel & Kjær 4191). Stimuli were presented with 10ms $\cos^2$ onset and offset ramps.

The vocalization-modulated noise and natural sound data used in this study are available for download at https://doi.org/10.5281/zenodo.3445557. A python library for loading data and fitting encoding models is available at https://github.com/LBHB/NEMS/.

Neurophysiological recording

After animals were prepared for experiments, we opened a small craniotomy over primary auditory cortex (A1). Extracellular neurophysiological activity was recorded using 1-4 independently positioned tungsten microelectrodes (FHC). Amplified (AM Systems) and digitized (National Instruments) signals were stored using MANTA open-source data acquisition software (Englitz et al., 2013). Recording sites were confirmed as being in A1 based on dorsal-ventral, high-to-low frequency tonotopy and relatively reliable and simple response properties (Atiani et al., 2014; Shamma et al., 1993). Some units may have been recorded from AAF, particularly in the high frequency region where tonotopic maps converge. Single units were sorted offline by bandpass filtering the raw trace (300-6000 Hz) and then applying PCA-based clustering algorithm

to spike-threshold events (S. V David et al., 2009). Neurons were considered isolated single units if standard deviation of spike amplitude was at least two times the noise floor, corresponding to > 95% isolation of spikes.

A pure-tone or broadband noise probe stimulus was played periodically to search for sound-activated neurons during electrode positioning. Upon unit isolation, a series of brief (100-ms duration, 100-ms interstimulus interval, 65 dB SPL) quarter-octave noise bursts was used to determine the range of frequencies that evoked a response and the best frequency (BF) that drove the strongest response. If a neuron did not respond to the noise bursts, the electrode was moved to a new recording depth. Thus our yield of 187/200 neurons responsive to vocalization-modulated noise overestimates the rate of responsiveness across the entire A1 population. Center frequencies of the vocalization-modulated noise stimuli were then selected based on this tuning curve, so that one or both of the noise bands fell in the frequency tuning curve measured with single noise bursts.

Tone detection task

Three ferrets were trained to perform a tone in noise detection task (Schwartz & David, 2018). The task used a go/no-go paradigm, in which animals were required to refrain from licking a water spout during presentation of vocalization-modulated noise until they heard the target tone (0.5 s duration, 0.1 s ramp) centered in one noise band at a random time (1, 1.5, 2, … or 5 s) after noise onset. To prevent timing strategies, the target time was distributed randomly with a flat hazard function (Heffner & Heffner, 1995). Target times varied across presentations of the same noise distractors so that animals could not use features in the noise to predict target onset.

In a block of behavioral trials, the target tone matched the center frequency and spatial position of one noise stream. Behavioral performance was quantified by hit rate (correct responses to targets vs. misses), false alarm rate (incorrect responses prior to the target), and a discrimination index (DI) that measured the area under the receiver operating characteristic (ROC) curve for hits and false alarms (Schwartz & David, 2018; Yin et al., 2010). A DI of 1.0

reflected perfect discriminability and 0.5 reflected chance performance. A detailed analysis of behavior is reported elsewhere (Schwartz & David, 2018). In the current study, only data from blocks with DI significantly greater than chance and correct trials were included in the analysis of neural encoding.

During recordings, one noise stream was centered over a recorded neuron's best frequency and the other was separated by 1-2 octaves. The target tone fell in only one stream on a single block of trials. Identical task stimuli were also presented during a passive condition, interleaved with behavioral blocks, during which period licking had no effect. Previous work compared activity between conditions when attention was directed into versus away from the neuron's receptive field (Schwartz & David, 2018). Because of the relatively small number of neurons showing both task-related and STP effects in the current study (see Figure 8), data were collapsed across the different target conditions. Instead, neural activity was compared for the vocalization-modulated noise stimuli between active and passive listening conditions.

Spectro-temporal receptive field models

Linear-nonlinear spectro-temporal receptive field (LN model). Vocalization-modulated noise was designed so that the random fluctuations in the two spectral channels could be used to measure spectro-temporal encoding properties. The LN model is a widely viewed as a current standard model for early stages of auditory processing (Aertsen & Johannesma, 1981; Machens et al., 2004; Radtke-Schuller et al., 2009; Theunissen et al., 2001). The LN model is an implementation of the generalized linear model (GLM), which is used widely across the auditory and other sensory systems (Calabrese et al., 2011; Paninski et al., 2004). In the first, linear stage of this model, a finite impulse response (FIR) filter, $h(x,u)$, is applied to the stimulus spectrogram, $s(x,t)$, to produce a linear prediction of time-varying spike rate, $r_L(t)$,

$$r_L(t) = \sum_{x=1}^{J} \sum_{u=0}^{U} h(x,u)s(x,t-u) \quad (1)$$

For the current study, the time lag of temporal integration, *u*, ranged from 0 to 150 ms. In the auditory system, this first, linear component of the LN model is commonly referred to as the spectro-temporal receptive field (STRF). In typical STRF analysis, the stimulus is broadband and variable across multiple spectral channels, *x*. Here, the stimulus spectrogram was composed of just two time-varying channels, and a simplified version of the linear filter was constructed in which x spanned just these two channels (i.e., J = 2), but we used larger values following spectral reweighting, below. A log compression was applied to the spectrogram to account for cochlear nonlinearities (offset 1 to force the compressed output to have nonnegative values, (Thorson et al., 2015)). Otherwise, this model functions as a traditional STRF.

In the second stage of the LN model, the output of the linear filter, $r_L(t)$ is transformed by a static, sigmoidal nonlinearity, which accounts for spike threshold and saturation. The current study used a double exponential sigmoid,

$$r(t) = b + A \exp\left[-\exp\left(\kappa(r_L(t) - b)\right)\right] \quad (2)$$

where r_0 is the baseline (spontaneous) spike rate, A is the maximum evoked rate, κ is the slope, and b is the baseline. The specific formulation of the output nonlinearity does not substantially impact relative performance of models in which other aspects of model architecture are manipulated (see Output nonlinearity controls, below).

Temporal filter parameterization and spectral reweighting. As models become more complex (i.e., require fitting more free parameters), they become more susceptible to estimation noise. While our fitting algorithm was designed to prevent overfitting to noise (see below), we found that constraining the temporal form of the linear filter in Eq. 1 improved performance over a model in which the filter was simple as set of weights for each time lag. Each spectral channel of the linear filter was constrained to be a damped oscillator, $h_{\mathrm{DO}}(x, u)$,

$$h_{\mathrm{DO}}(x, u) = G \exp(-\tau|u - u_0|^+) \sin(f|u - u_0|^+), \quad (3)$$

Requiring four free parameters: gain, G; latency, u_0; duration, τ; and modulation frequency, f.

Because the damped oscillator constrains temporal tuning, we then considered the possibility that more than $J = 2$ spectral channels might be optimal for explaining neural responses to the two band vocalization-modulated noise stimulus. To allow for more than two spectral channels, we defined a reweighted stimulus, $s_R(j,t)$, computed as the input stimulus scaled by coefficients, $w(i,j)$,

$$s_R(j,t) = \sum_{i=1}^{2} w(i,j)s(i,t) \tag{4}$$

where $j = 1 \ldots J$ maps the stimulus to a J-dimensional space. This reweighted stimulus provides input to the a damped oscillator, now also with J channels,

$$r_L(t) = \sum_{x=1}^{J} \sum_{u=0}^{U} h_{\mathrm{DO}}(x,u)s_R(x,t-u) \tag{5}$$

For the current study (except for controls, see below), the output of Eq. 5 is then transformed by the output nonlinearity (Eq. 2) to produce a predicted time-varying spike rate.

While we describe the LN model as a sequence of linear transformations—spectral filtering followed by temporal filtering—these two stages can be combined into a single linear spectro-temporal filter. We describe them as separate stages to frame the local STP model, below, where nonlinear adaptation is inserted between the two linear filtering stages.

Local short-term plasticity (STP) model. As several studies have demonstrated, the LN model captures import aspects of spectro-temporal coding but fails to account completely for time-varying sound evoked activity in auditory cortex (Atencio et al., 2008; Machens et al., 2004; Rabinowitz et al., 2012; Williamson et al., 2016). In particular, the LN model fails to account for the temporal dynamics of sound-evoked activity (S. V David et al., 2009; S. V David & Shamma, 2013). Short-term synaptic plasticity (STP), the depression or facilitation of synaptic efficacy following repeated activation, has been proposed as one mechanism for nonlinear dynamics in neural networks (Angeloni & Geffen, 2018; Tsodyks et al., 1998). A previous study showed that

an LN model for A1 that incorporated STP was able to better explain the dynamics of responses to a single noise band with natural temporal modulations (S. V David & Shamma, 2013). However, because that study utilized vocalization-modulated noise comprised of only a single noise band, it was not clear whether the nonlinear adaptation was global, affecting responses to all stimuli equally, or local, affecting only a subset of inputs independently. Because the current study used multiple noise channels, it could compare a global STP model, in which adaptation affected all input channels, to a local STP model, in which adaptation was spectrally tuned and could affect just a subset of inputs.

The effects of nonlinear adaptation were captured with a simple, two-parameter model of STP (Tsodyks et al., 1998),

$$d(i,t) = d(i,t-1) + s_R(i,t-1)[1 - d(i,t-1)]v_i - \frac{d(i,t-1)}{\tau_i} \qquad (6)$$

where $d(i,t)$ describes the change in gain for stimulus channel i at time t. The change in available synaptic resources (release probability), v_i, captures the strength of plasticity, and the recovery time constant, τ_i, determines how quickly the plasticity returns to baseline. Values of $d < 1$ correspond to depression (driven by $v_i > 0$) and $d > 1$ correspond to facilitation ($v_i < 0$). In the local STP model, each input channel of the stimulus is scaled by $d(i,t)$ computed for that channel,

$$s_{STP}(i,t) = d(i,t)s_R(i,t) \qquad (7)$$

This nonlinearly filtered stimulus is then provided as input to the LN filter (Eqs. 5, 2) to predict the time-varying response. Note that if the strength of STP is 0 (i.e., $v_i = 0$), then the STP model reduces to the LN model.

We also note that the local STP model uses the reweighted stimulus as its input. The reweighting allows the model to account for adaptation at multiple timescales on inputs from the same spectral band. Although the input is comprised of just two channels, the subsequent nonlinear filtering means that allowing the reweighted stimulus channel count, J, or rank, to be greater than two can increase model predictive power. In the current study, we evaluated models

with rank J = 1-5. Predictive power was highest for J = 5. Higher values of J could, in theory, produce even better performance, but we did not observe further improvements for the current dataset.

Global STP model. We considered two control models to test for the specific benefit of spectrally tuned adaptation on model performance. One possible alternative is that a single, global adaptation is able to account for nonlinear temporal dynamics. To model global adaptation, the global STP model applied STP to the output of the linear filter (Eq. 5) before applying the static nonlinearity (Eq. 2). Thus, a single adaptation term was applied to all incoming stimuli, rather than allowing for the channel-specific adaptation in the local STP model. There is no simple biophysical interpretation of the global STP mechanism, but it can be thought of as a postsynaptic effect, capturing nonlinear dynamics similar to STP, but after integration across spectral channels. We compared performance of this model to a variant in which stimulus gain is averaged across spectral channels before scaling the input stimulus,

$$\bar{d}(t) = \langle d(i,t) \rangle_i \tag{8}$$

We found no difference between this common input STP model and the global STP model. Because the global model required fewer free parameters, we focused on this model for the comparisons in this study.

Local rectification model. Although spectral reweighting can improve the performance of the LN model by increasing the rank of the linear filter, it can still only account for linear transformations of the input stimulus. The STP model could, in theory, benefit simply from the fact that reweighted spectral inputs undergo any nonlinear transformation prior to the temporal filter. To control for the possibility that the STP nonlinearity is not specifically beneficial to model performance, we developed a local rectification model, in which the reweighted spectral inputs were linearly rectified with threshold s_0 prior to temporal filtering,

$$s_+(j,t) = |s_R(j,t) - s_0(j)|^+ \quad (9)$$

The rectified reweighted stimulus then provided the input to the LN model specified in Eqs. 7 and 2.

The set of encoding models described above represents a hierarchy of model architectures with increasing complexity, in that each successive model requires additional free parameters. Each model can be cast as a sequence of transformations applied to the stimulus, and the output of the final transformation is the predicted time-varying response (Figure 3).

Fit procedure. Spike rate data and stimulus spectrograms were binned at 10 ms before analysis (no smoothing). The entire parameter set was fit separately for each model architecture. Data preprocessing, model fitting, and model validation were performed using the NEMS library in Python (Pennington & David, 2022). Identical estimation data from each neuron and the same gradient descent algorithm were used for each model (L-BFGS-B, (Byrd et al., 1995)). The optimization minimized mean squared error (MSE) with shrinkage, a form of early stopping in which the standard MSE value is scaled by its standard error (Thorson et al., 2015). The use of parameterized temporal filters provided an effective regularization, as it constrained the shape of the temporal filter to be smooth and sinusoidal. Models were initialized at 10 random initial conditions (except for local minimum controls, see below), and the final model was selected as the one that produced the lowest MSE with shrinkage for the estimation data. Scripts demonstrating model fits using the NEMS library are available with the data at https://doi.org/10.5281/zenodo.3445557.

The ability of the encoding model to describe a neuron's function was assessed by measuring the accuracy with which it predicted time varying activity in a held-out validation dataset that was not used for model estimation. The prediction correlation was computed as the correlation coefficient (Pearson's *R*) between the predicted and actual PSTH response. Raw correlation scores were corrected to account for sampling limitations that produce noise in the actual

response (Hsu et al., 2004). A prediction correlation of R=1 indicated perfect prediction accuracy, and a value of R=0 indicated chance performance. All models were fit and tested using the same estimation and validation data sets. Significant differences in prediction accuracy across the neural population were determined by a Wilcoxon sign test.

In a previous study involving just a single stream of vocalization-modulated noise, we tested our fitting procedure on simulated data produced by either an LN model or STP model. The estimated models captured the presence or absence of the STP nonlinearity accurately (S. V David & Shamma, 2013). In addition, the simulations revealed that LN models could capture some aspects of the nonlinear adapting data, but estimated temporal filter properties did not match the actual temporal filter properties.

For data from the behavior experiments, which was all fit using single trials, 10-fold cross validation was used, on top of the procedure described above. Ten interleaved, non-overlapping validation subsets were drawn from the entire passive plus active data. The above fit algorithm was then applied to corresponding 90% estimation set, and the resulting model was used to predict the validation subset. Prediction accuracy was assessed for the conjunction of the 10 validation sets. Model parameters were largely consistent across estimation sets and average fit values are reported in the Results.

Output nonlinearity control. In a previous study, we compared performance of a variety of different static nonlinearities for LN models and found that the double exponential sigmoid (Eq. 2) performed slightly, but consistently, better than other formulations of the output nonlinearity for A1 encoding models fit using natural vocalization stimuli (Thorson et al., 2015). We performed a similar comparison using the speech-modulated vocalization data, comparing LN and local STP models with four different output functions (Figure 10 A): linear pass-through,

$$r(t) = r_L(t); \tag{10}$$

linear rectification,

$$r(t) = |r_L(t) - b|^+ + r_0, \quad (11)$$

with threshold *b* and spontaneous rate r_0; logistic sigmoid (Fitzgerald et al., 2011; Rabinowitz et al., 2012),

$$r(t) = r_0 + \frac{A}{1+\exp[-(r_L(t)-b)/\kappa]}, \quad (12)$$

and the double exponential sigmoid (Eq. 2). As in the previous study, the double exponential sigmoid performed best for the LN model. The STP model incorporating a given output function always performed better than the corresponding LN model with the same output function, and the double exponential performed best overall. Thus for the rest of the study, we focused on models using the double exponential nonlinearity.

Temporal parameterization control. The use of a damped oscillator to constrain model temporal dynamics could, in theory, be suboptimal for describing neural response dynamics. We compared performance of the damped oscillator models to LN and local STP models with nonparametric temporal filters, that is, where the temporal filter was simply a vector of weights convolved with the stimulus at each time point (Figure 10 B). For the LN model, the parameterization had no significant impact on average model performance ($p > 0.05$, sign test). For the local STP model, performance was higher for the parameterized model ($p < 10\text{-}5$, sign test), indicating that the parameterization was an effective form of regularization.

Local minimum control. While it has been shown that linear filters are well-behaved (*i.e.*, convex) and thus not subject to problems of local minima during fitting, it is more difficult to determine if local minima are adversely affecting performance of nonlinear or parametric models, such as the STP and damped oscillator, respectively, used in the current study. To determine if these models were negatively impacted by local minima during fitting, we compared performance of models fit from a single initial condition to the best model (determined using only estimation data) starting from 10 random initial conditions. Performance was compared for 4 different architectures: LN and local STP models, each with parametric (damped oscillator) or

nonparametric temporal filters (Figure 10 B). Each model was tested with the validation data. For the LN models, we saw no significant effect of using multiple initial conditions. For the local STP model, we saw a small but significant improvement when multiple initial conditions were used. Thus for the majority of results presented here, models were fit using 10 random initial conditions.

Stimulus specific adaptation analysis

Sound-evoked activity recorded during presentation of the oddball noise burst sequences was modeled using the LN model, global STP model, and local STP model, as described above. To assess stimulus specific adaptation (SSA) , an SSA index (SI) was used to measure the relative enhancement of responses to oddball versus standard noise bursts (Pérez-González & Malmierca, 2014; Ulanovsky et al., 2003).

$$SI = \frac{\bar{r}_{\text{odd}} - \bar{r}_{\text{std}}}{\bar{r}_{\text{odd}} + \bar{r}_{\text{std}}} \qquad (13)$$

Here $\bar{r}_{\text{odd}}$ and $\bar{r}_{\text{std}}$ are the average response across bursts of both center frequencies, in the oddball and standard conditions, respectively. Neuronal response was calculated as the integral over time of the PSTH during the sound presentation. Significance SI was calculated with a shuffle test in which the identity of tones (oddball or standard) was randomly swapped. To determine how well each model could account for SSA, SI was calculated for model predictions, also using Eq. 13. We then assessed the accuracy of SI predicted by models in two ways: First we computed the correlation coefficient between actual and predicted SI for all the recorded cells that showed significant SI. Second, as the population mean of the squared difference between the actual and predicted SI calculated individually for each cell.

Behavior-dependent encoding models.

To measure effects of behavioral state on spectro-temporal coding, we estimated behavior-dependent models, by allowing some or all of the fit parameters to vary between passive and active behavioral conditions (Schwartz & David, 2018). Having established the efficacy of the reweighted STP model for passive-listening data, analysis focused on this architecture for the

behavioral data. First, a behavior-independent model provided a baseline, for which all model parameters were fixed across behavior conditions. Second, a behavior-dependent static nonlinearity allowed parameters of the static nonlinearity (Eq. 2) to vary between behavior states but kept all other parameters fixed between conditions. Third, both the linear filter parameters and static nonlinearity (Eqs. 5, 2) were allowed to vary between behavior conditions, with reweighting and STP parameters fixed across conditions. Finally, all model parameters were allowed to vary between behavior conditions. Thus, this progression of models explored the benefit of allowing increased influence of changes in behavioral state on spectro-temporal coding.

Behavior-dependent models were fit using a sequential gradient descent algorithm in the NEMS library. All models were initially fit using a behavior-independent model. The specified behavior-dependent parameters were then allowed to vary between behavioral states in a subsequent application of gradient descent. Model performance was compared as for the passive-listening data described above. For each neuron, prediction accuracy was assessed using a validation set drawn from both active and passive conditions, which was excluded from fitting, and was always the same across all models. Significant behavioral effects were indicated by improved prediction correlation for behavior-dependent models over the behavior-independent model. Changes in tuning were measured by comparing model fit parameters between behavior conditions.

Nonlinear encoding models for natural sounds

Encoding of natural sounds was modeled using a similar approach as for the vocalization-modulated noise. Here we focused on two models, a baseline LN model and a local STP model (Figure 9). Because natural sounds contain spectral features that vary across a large number of spectral channels, a different spectral filtering process was required prior to the STP stage. This was achieved using a reduced-rank model, where the full spectro-temporal filter in the linear stage was computed from the product of a small number of spectral and temporal filters (Simon et al., 2007; Thorson et al., 2015). The input spectrogram was computed from a bank of log-spaced

gammatone filters, $s(i,t)$, with N = 18 spectral channels (Katsiamis et al., 2007). Spectral tuning was modeled with a bank of J weight vectors, $w(i,j)$, each of which computed a linear weighted sum of the log-compressed input spectrogram,

$$s_N(j,t) = \sum_{i=1}^{N} w(i,j)s(i,t) \tag{14}$$

The reweighted stimuli, $s_N(i,j)$, were provided as inputs to the LN and STP models (see above). Each spectral filter was initialized to have constant weights across channels. Model fitting and testing were performed using the same procedures as for the vocalization-modulated noise data (see above).

Statistical methods

To test whether the prediction of a model for a single neuron was significantly better than chance (i.e., the model could account for any auditory response), we performed a permutation test. The predicted response was shuffled across time 1000 times, and the prediction correlation was calculated for each shuffle. The distribution of shuffled correlations defined a noise floor, and a p value was defined as the fraction of shuffled correlations greater than the correlation for the actual prediction. The Bonferroni method was used to correct for multiple comparisons when assessing significance across any of multiple models.

To compare the performance of two models for a single neuron, we used a jackknifed t-test. The Pearson's correlation coefficient between the actual response and response predicted by each model was calculated for 20 jackknife resamples. We then calculated the mean and standard error on the mean from the jackknifed measures (Efron & Tibshirani, 1986). The prediction of two models was considered significantly different at $p<0.05$ if the difference of the means was greater than the sum of the standard errors.

To test whether the calculated SSA Index (SI) was significantly different than chance, we performed a permutation test in which the identity of tones (standards, oddball) was shuffled, and the SI was calculated 1000 times. The real SI value was then compared to the noise floor

distributions. Finally, for comparing model performance across collections of neurons, we performed a Wilcoxon signed rank test (sign test) between the median prediction correlation across neurons for each model.

Acknowledgements

The authors would like to thank Henry Cooney, Sean Slee, and Daniela Saderi for assistance with behavioral training and neurophysiological recording.

Chapter 3. Sensory context for coding of natural sounds in auditory cortex

Abstract

Discriminating sound objects lasting seconds or fractions of a second requires integrating information over these time scales. Prior studies have shown the extent of this integration in single cells, and some of the synaptic mechanisms responsible for it. Here we explore how these neuronal responses are coordinated from the perspective of circuits and neuronal populations. We used the difference in response to a single probe sound after two different contextualizing sounds as a proxy to the memory and temporal integration of a neuron. We found context dependent differences that lasted past the temporal window of a traditional spectrotemporal receptive field. Individual neurons showed contextual effects restricted to specific stimuli combinations. However, different neurons within a population showed different stimuli preferences, thus forming a sparse code that covered more of the stimulus space than any constituent neuron. Through modeling, we posit neuron specific spectro-temporal tuning, alongside network connections, as the underlying mechanisms forming this sparse long-lasting representation of past context.

Introduction

Natural sounds are characterized by diverse temporal dynamics, like amplitude modulation at different rates. In behaviorally relevant sounds like speech, these dynamics span a range of timescales, from tens of milliseconds for phonemes, to hundreds of milliseconds for syllables and longer times for words, phrases and so on (Chomsky & Halle, 1968). Keeping track of temporal information over these diverse timescales is critical for computation and discrimination of important sound features (Norman-Haignere et al., 2022). Hearing impairment can impair temporal integration, and this deficit is likely to impair speech comprehension (Albouy et al., 2020).

Neurons in the auditory cortex respond to specific spectrotemporal features of ongoing sound. Prior studies have characterized auditory neuronal tuning with a linear model called the spectro-temporal receptive field (STRF) (Aertsen & Johannesma, 1981). These models can account for temporal integration up to about 150 ms (Atiani et al., 2014), and therefore accurately capture the response of neurons in early stages of the auditory pathway, which linearly respond to sound and quickly track its changes over time (Aertsen & Johannesma, 1981; Escabí & Read, 2003; Kowalski et al., 1996). However, these models cannot describe several known nonlinear response properties in auditory cortex (Atiani et al., 2014; Christianson et al., 2008), including nonlinear coding of modulation rate (Joris et al., 2004; T. Lu et al., 2001; Sharpee et al., 2011), adaptations to particular parts of the stimuli (Ulanovsky et al., 2003), invariance to background noise (Rabinowitz et al., 2013) and other sound features that might last hundreds of milliseconds (Sharpee et al., 2011), likely to be nonlinear functions across longer timescales.

Model-free analysis has shown that neurons in the auditory cortex have memories longer than those described by STRFs. Both subthreshold potentials (Asari & Zador, 2009) as well as spikes (Asokan et al., 2021) can hold memory of prior stimuli lasting more than 1 second. One well-known integration property is the reduced neuronal response to repeated but not novel sounds, denominated stimulus specific adaptation or SSA (Ulanovsky et al., 2003). These studies have focused on simplified sounds with controlled parameters, which might not capture the more complex interaction associated with natural sounds (Theunissen et al., 2000).

Several neuronal and synaptic mechanisms that can contribute to temporal integration have been elucidated (Silver, 2010). This neuronal integration is likely to be amplified and modulated by circuit and population dynamics (Buonomano & Maass, 2009), thus extending the computational and representational capabilities of a whole population of neurons. Recordings from large populations of neurons shed light on sensory representation at this larger scale (Du et al., 2011; Steinmetz et al., 2021). Among other phenomena, this approach demonstrates the existence of sparse codes in touch (Lyall et al., 2021), the dimensionality of representations in

vision (Stringer, Pachitariu, Steinmetz, Carandini, et al., 2019), and a distributed encoding of time (Runyan et al., 2017). The characteristics of temporal integration in population of neurons across different auditory cortex regions, and the strategies taken by these populations to represent natural stimuli remain an open question.

To gain a better insight into the mechanisms underlying temporal integration, we used linear microelectrode arrays to record the activity of multiple neurons in auditory cortex and quantify the influence of recent stimuli on the response to ongoing natural sounds. We observed effects of sensory context (temporal integration) lasting up to several hundred milliseconds in primary and secondary fields, with a tendency toward stronger and longer-lasting effects in secondary fields. Individual neurons tended to be sensitive to a small number of contexts, but the aggregate population activity formed a sparse representation tiling a much larger space of sensory context. Using encoding model analysis, we determined that local population dynamics can account for these long-lasting contextual effects, which cannot be explained by a traditional STRF model

Results

Responses of neurons in auditory cortex to natural sounds are modulated by sensory context

To measure the effects of sensory context on neural coding of sound, we recorded single-unit neural activity in auditory cortex (AC) of awake, passively listening ferrets during the presentation of sequences of 1-s natural sound samples. Activity was recorded from neurons in primary auditory cortex (A1) and a secondary auditory cortical field (peri-ectosylvian gyrus, PEG). Sounds were presented repeatedly and in varying order, so that the neural response to the same probe sound was recorded following many different contexts, defined as the immediately preceding sound (Figure 11 A). Neural activity was recorded from linear microelectrode arrays so that the activity of tens of single units were collected simultaneously (Figure 11 B, 1724 units, 64 recording sites, 5 animals).

To measure the contextual integration window of an auditory cortex (AC) neuron, we computed the difference in response to a probe sound following different context sounds (Figure 11 D,H). A contextual modulation profile was calculated as the timewise difference between probe PSTHs following two different contexts. We computed a T-score for each time bin (Figure 11 F,J). Significant differences across multiple consecutive timepoints were identified using a cluster mass method (Maris & Oostenveld, 2007). We performed this comparison for each contextual instance, defined as a pair of contexts preceding a probe.

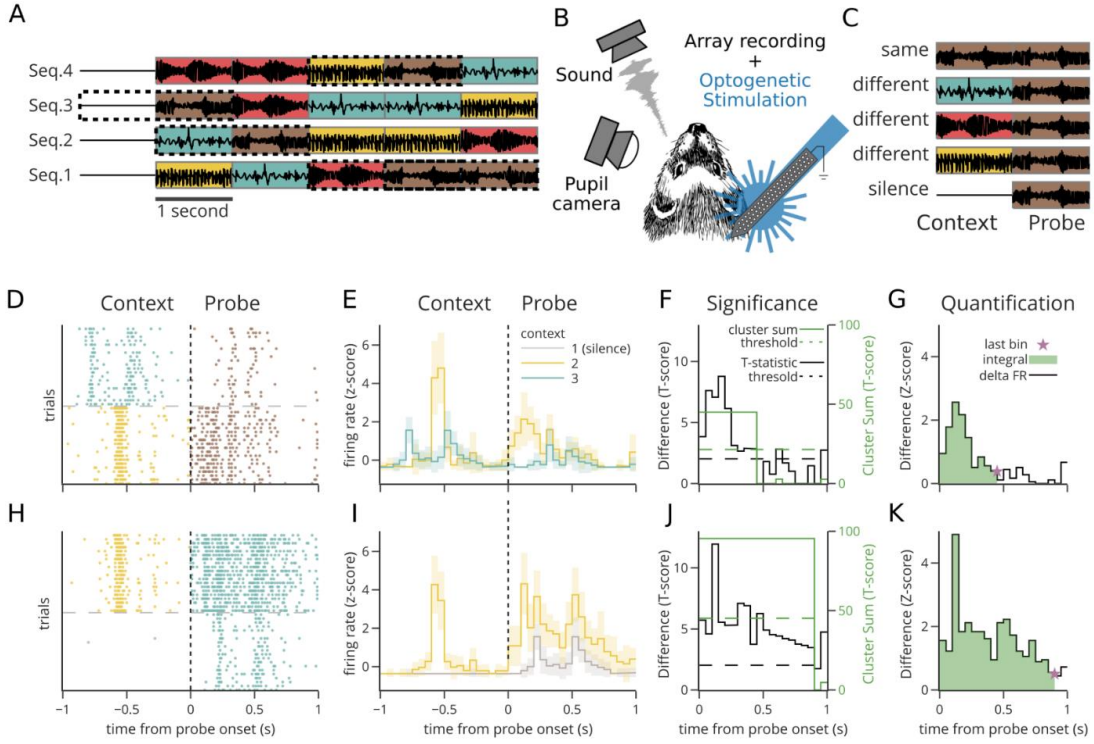

Figure 11. Effect of preceding sensory context on the response to a probe stimulus.
A. Example sequences composed from four 1-second natural sounds. Sounds were ordered such that every different sound (indicated by color) followed every other sound, silence, and itself exactly once **B.** Sounds were played to passive listening ferrets while recording from the auditory cortex with a multi electrode array. Pupillometry and photo-tagging of inhibitory interneurons were also performed. **C.** For analysis we considered the response to one probe (brown, dotted boxes on a.) after all contexts: different sounds, silence, or the same sound. **D.** Raster of an example neuron response to multiple repetitions of the same probe (brown) after two different contexts (teal and yellow, 20 trials per context). **E.** Trial average response (PSTH) of the data in (C.) showing the mean (line) and SEM (shading). Line color indicates the context stimulus. **F.** Quantification of context effects: a T-score between the probe response after two contexts was calculated for every time bin (solid black line). Clustered T-scores over a threshold (dashed black line, α=0.05) were summed (solid green line). The significance of cluster scores was determined through a shuffle test (dashed green line, α=0.05). **G.** The magnitude of significant contextual effects (Δ firing rate, black solid line) was quantified as its amplitude (green area under the curve, 0.559 Z-score*s) and duration (last significant bin, purple star, 450ms) **H,I,J,K.**

Same as D,E,F,G. For a different set of two contexts and one probe for the same neuron as in (d.) (integral: 1.508 Z-score*s, last bin: 900ms).

Neurons were presented with 50 or 550 context-probe instances, depending on the number of distinct sounds presented (N=4 or 10 distinct sounds, respectively). We recorded the activity of 1724 AC neurons, yielding a total of 502537 combinations of context pair, probe, and neuron. In total, only 9.135% (*n*=45905) of all these contextual instances showed significant modulation. However, 71.52% of all neurons showed significant effects for at least one contextual instance (*n*=1233/1724, $p<0.05$, multiple comparisons correction). Because effects were highly variable within and across neurons, we analyzed each contextual instance as a distinct data point.

Amplitude and duration of contextual modulation varies across neurons.

Contextual modulation profiles tended to be strongest immediately following probe onset and then decay over time (Figure 11 E, H). These dynamics are consistent with the idea of a finite integration window (Asari & Zador, 2009; Atiani et al., 2014; Norman-Haignere et al., 2022). However, the time-course showed great diversity across contextual instances and could be complex. For example, some modulation profiles had multiple peaks and valleys (Figure 11 F, J). We therefore used a non-parametric approach to quantify their amplitude and duration. Amplitude was defined as the integral of the absolute delta firing rate ($\int\Delta$ Z-score), across the probe response time bins with a significant difference between contexts, as identified with the T-score cluster mass test. Duration was defined as the last significant bin (Figure 11 G, K).

The amplitude and duration of contextual effects were distributed unimodally and were correlated with across the neural population (Figure 12 A. r=0.479, p=0, Pearson's correlation). Some contextual effects lasted only briefly after probe onset and therefore had relatively small amplitude, while long lasting effects generally had greater overall amplitude. However, there were many examples of long lasting, low amplitude effects due to late onset of the contextual effect. Across all significant contextual instances in AC (A1 and PEG), both duration and amplitude were highly variable (mean±std, duration: 249.25±208.70ms. Amplitude: 0.236±0.189 Z-score*s). In

many instances (n=667, 1.4% of all significant instances), the contextual modulation spanned the entire probe duration (1s), suggesting that contextual effects can last seconds, consistent with previous reports of AC integration windows in other preparations (Asokan et al., 2021; Norman-Haignere et al., 2022).

Context effects are stronger and longer-lasting in secondary auditory cortex

Compared to A1, neurons in PEG have complex receptive fields, associated with longer response latencies and longer integration windows (Atiani et al., 2014; Bizley et al., 2005; Norman-Haignere et al., 2022). Consistent with these previous observations, PEG neurons showed longer-lasting contextual effects than A1 (Duration, mean±SEM. A1: 244.97±1.36, PEG: 254.25±1.40ms. Figure 12 E). In addition, we observed a striking difference in the amplitude of the contextual effects (mean±SEM. A1: 0.23±0.001, PEG: 0.25±0.001 Z-score*s, Figure 12 C). This result suggests that the relative weight given to representing past memory versus current stimuli differs between areas, in addition to the duration of the integration window.

Magnitude of contextual effects depends on context category

We speculated that short term stimulus-specific adaptation could contribute to contextual effects. When the preceding context is silence, a probe would be a novel stimulus to which responses are not adapted and are therefore salient (onset response). In contrast, a probe that is a repeat of the context sound would be one for which responses are already adapted, similarly to what is observed in the case of stimulus-specific adaptation (Carbajal & Malmierca, 2018; Ulanovsky et al., 2003). To explore this possibility, we compared contextual effects after grouping by the relationship of context to the probe: Silence, the Same sound as the probe, or a Different sound than the probe (Figure 11 B). We used multivariate regression to quantify the contribution of the different context categories to the probe response. According to this model, context effect amplitude and duration were each a weighted sum of the categories comprising a contextual instance. Amplitude and duration were highest when Silence was one of the contexts being

compared (Amplitude: +21.43%, p<0.001; Duration: +20.58%, p<0.001. Change relative to Different, T-test). Effects were weakest when one context was the *Same* sound as the probe (Amplitude: -3.26%, p<0.001; Duration: 1.31% p=0.176. Change relative to Different, T-test). When one of the contexts was a Different sound, the effect fell between the extremes of Silence and Same sound. Differences in duration of context effects were less systematic than for amplitude. Duration did not change between Silence and Different conditions. Since the distribution of amplitude and duration were highly non normal, we validated the results from the multivariate regression with a nonparametric ANOVA and post hoc test, which confirmed the regression results (Figure 12 B, D. integral mean±SEM. same: 0.23±0.0018, diff: 0.24±0.0009, silence: 0.27±0.0021 Z-scores; last_bin mean±SEM. same: 246.92±2.16, diff: 247.83±0.98, silence: 285.79±2.08 ms). This pattern of greater modulation following sharper transitions is consistent with the possibility that adaptation to the spectro-temporal features of the context sound contributes to contextual effects.

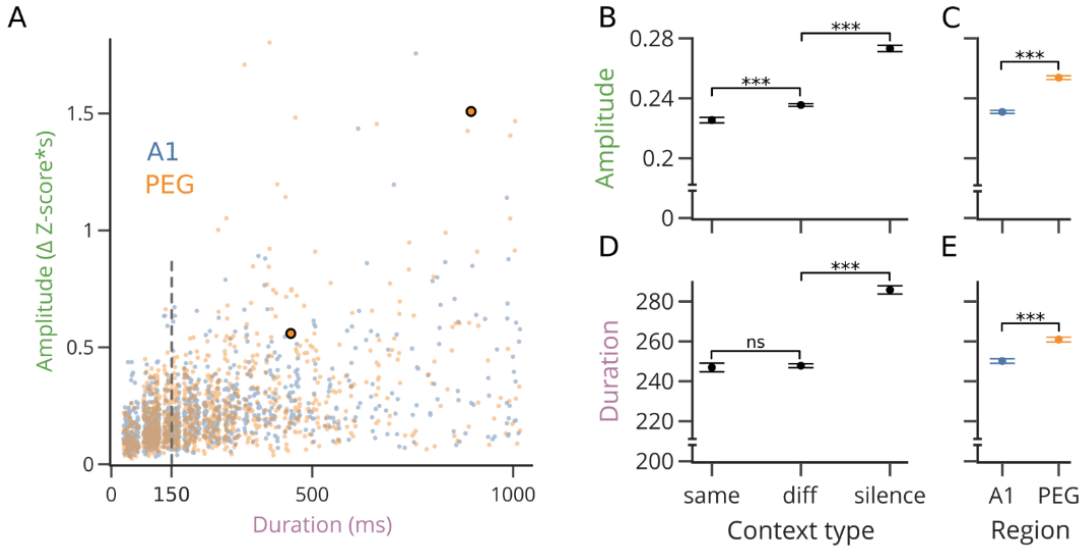

Figure 12. Context effects magnitude across cortical regions and context types.
A. Distribution of amplitude and duration of contextual effects. Each dot indicates magnitude and duration for one contextual instance—a combination of a neuron, a context-pair, and a probe. The two example instances from fig.1 are highlighted. Dot color indicates data from primary auditory cortex (A1, blue) and a secondary region of auditory cortex (peri ectosylvian gyrus, PEG, orange). For display clarity the duration values have been jittered (actual values are discrete in the 20Hz sampling rate) and the data was decimated by taking a random subset of 1000 instances per brain region (A1: n=24711 instances,

n=709 neurons. PEG: n=21195 instances, n=523 neurons). The dashed gray line indicates the standard temporal integration window of standard LN STRFs. **B.** Mean contextual effect amplitude as a function of one of the contexts being silence or a sound equal (same) or different (diff) than the probe (mean and SEM, Kruskal Wallis p<0.001. Dunn post hoc with Bonferroni correction, all p<0.001). **C.** Difference in mean context effect amplitude between brain regions, plotted as in (b.) (Kruskal Wallis p<0.001) **D,E.** Same as B,C but for contextual effect duration (Context type: Kruskal Wallis p<0.01, Dunn post hoc with Bonferroni correction: diff-vs-same p=0.563, diff-vs-silence p<0.001, silence-vs-same p<0.001. Region: Kruskal Wallis p<0.001)

Sparse representation of context

We exposed each neuron to multiple combinations of pairs of contexts and probes, which we defined as the context space (40 and 550 combinations for datasets composed of 4 and 10 different natural sound samples, respectively). Not all sound combinations elicited significant contextual effects; therefore, individual neurons were modulated in a limited extent of the context space (Figure 13 A, 4-sound examples). On average, a neuron covered only 11.2 ± 0.374% (Mean ± SEM) of the space (Figure 13 C). However, different neurons from the same recording showed effects in different regions of context space. We thus hypothesize that neurons belonging to the same local population represent context using a sparse code (Olshausen & Field, 2004).

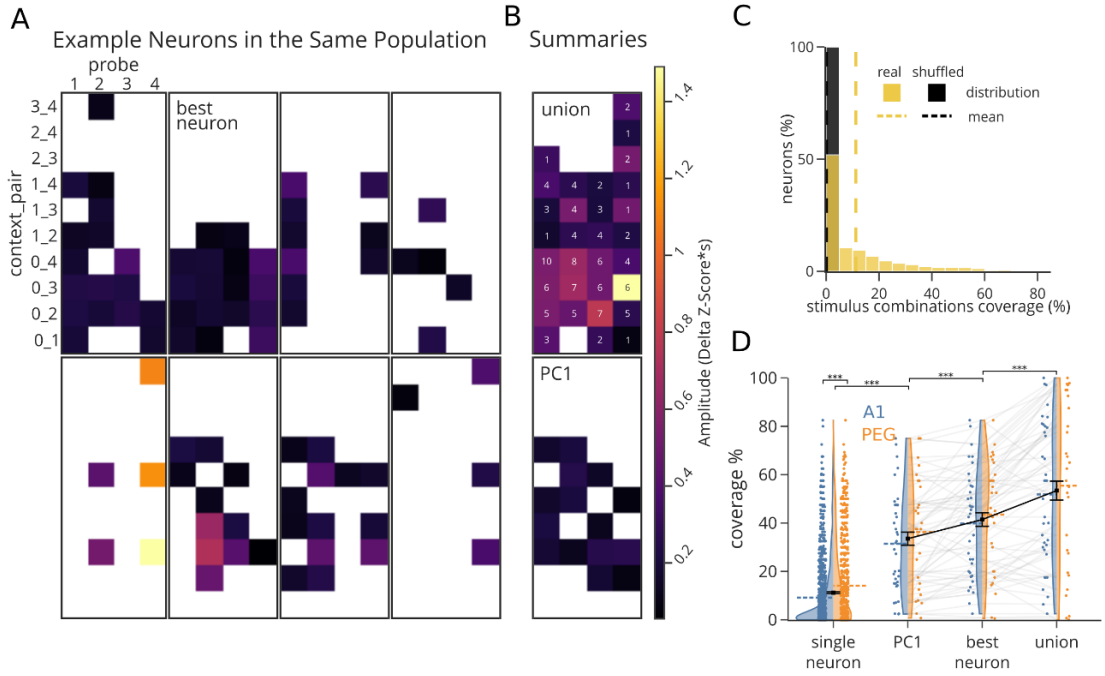

Figure 13. Sparse population code for contextual effects.
A. Contextual coverage, illustrated with the amplitude of contextual effects for all instances of context pairs and probes (4 distinct probes and 10 context pairs) for 8 example neurons recorded simultaneously.

Y-axis pairs indicate indices of sounds defining the context pair, with 0 denoting silence as a context. Only significant contextual effects are colored. The number of significant instances varies from 5 to 17 across neurons. **B.** Union of contextual effects across all neurons in the recording site (top, number denotes total of neurons showing significant effects per stimulus combination with further Bonferroni correction for number of neurons) and computed from the first principal component (PC1) of the population activity (bottom) **C.** Distribution of percent significant context instances across recorded neurons (yellow, n=1722, mean=11.20) and a null distribution obtained by randomly permuting trial context (black, mean=0.04) **D.** Distribution of percent contextual coverage for single neurons (A1 n=1006, PEG n=716 neurons), and neurons pooled across each recording site (A1 n=36, PEG n=28 sites. Blue: A1, Orange: PEG. Dashed colored lines indicate the region mean. Black square and error bars show mean and SEM for data pooled across regions). All population summaries showed greater contextual coverage than single neurons ($p<0.001$, Wilcoxon rank sum test, Bonferroni corrected). The best neuron from each site showed greater contextual coverage than PC1, but less than the union (PC1 vs best neuron: $p<0.001$, best neuron vs union: $p<0.001$. Wilcoxon signed-rank test, Bonferroni correction). PEG showed greater contextual coverage relative to A1 for individual neurons, but not for the population summaries (single neuron: $p<0.001$, mean A1=9.1, PEG=14.04; PC1: $p=0.21$, mean A1=31.43, PEG=36.24; best neuron: $p=0.23$, mean A1=39.84, PEG=43.50; union: $p=0.32$, mean A1=51.80, PEG=55.46. Wilcoxon rank sum test, Bonferroni correction).

It has been proposed that sparse codes are widely prevalent across sensory systems (Lyall et al., 2021; Olshausen & Field, 1996; Zhang et al., 2019). Sparse codes provide multiple advantages in associative learning, storage capacity, energy efficiency and facility to read out the encoded information (Beyeler et al., 2019; Olshausen & Field, 2004). To be useful for guiding behavior, sensory context must be read out by downstream neurons. We constructed two hypothetical decoders. A generalized decoder utilizes the pooled activity of a local population (first principal component, Figure 13 B bottom), and a specialized decoder takes into account the modulations distinct to each neuron (Union, Figure 13 B top).

The generalized decoder showed greater coverage of contextual space than the average single neuron. However, in most cases it was outperformed by the *best neuron* in the population, i.e., the single neuron with the greatest contextual coverage. Moreover, the Union, which can be thought of as an optimal context decoder, showed a significantly greater coverage of contextual space than both the first PC and the best neuron in population (Figure 13 E). The greater coverage by the Union is consistent with a sparse code, in which individual neurons each are modulated by a small number of contexts, but their joint activity provides information about a wide range of contexts.

Context effects are weaker but more common in putative inhibitory interneurons

Previous work has implicated inhibitory interneurons as having specialized roles in temporal processing of sound (Wehr & Zador, 2003), SSA (Natan et al., 2015; Yarden et al., 2022) and in sensory integration (Studer & Barkat, 2022). Thus, we hypothesized that they also play a distinct role in the representation of sensory context. We used a viral approach to express Channelrhodopsin (ChR2) selectively in inhibitory interneurons (Dimidschstein et al., 2016), which were then identified by optotagging (Figure 14 A). Putative inhibitory interneurons and pyramidal cells are also distinguished by the width of their spike wave form (Figure 14 B, C, (Trainito et al., 2019)). The narrow spike shapes of the optotagged neurons were consistent with that of putative inhibitory interneurons.

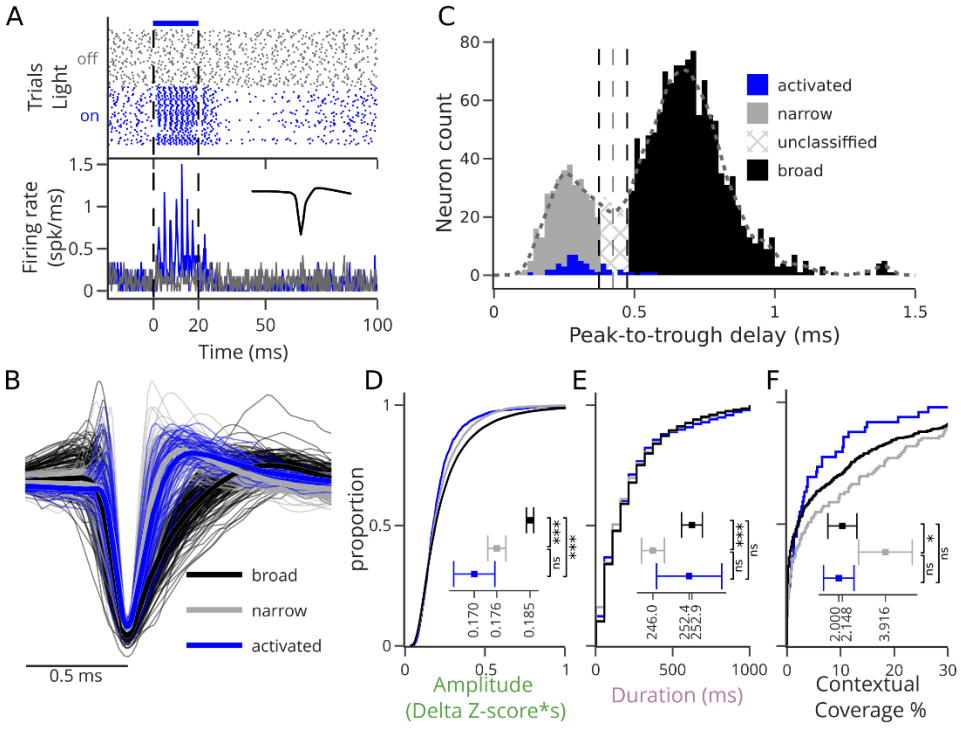

Figure 14 Context effects are weaker but more common in inhibitory interneurons.
A. Example single trial responses (top) and average PSTH (bottom) of a photo-tagged neuron to silence (gray) or a continuous 20ms light flashe (blue). Vertical dashed lines show light onset and offset. The inset in the lower panel shows the average spike waveform. A neuron was classified as optotagged if it responded with sustained spiking, starting <5ms after the light onset, and showing a reliable spiking pattern between trials. B. Example mean waveforms (thin lines) and average (thick lines) for neurons classified as narrow spiking (gray), broad spiking (black), and photoactivated neurons (blue). Waveforms normalized to a fixed peak. For clarity, only 500 random individual examples are shown per color. C. Histogram of spike peak-to-trough delay colored by cell type: narrow-spiking (n=301 neurons; peak-to-

trough delay<0.37ms; putative inhibitory; gray), broad-spiking (n=1172 neurons, peak-to-trough delay>0.47ms; putative excitatory; black) and photoactivated inhibitory interneurons (n=51 neurons, blue). Single units with intermediate peak-to-trough values were unclassified. D,E,F. Cumulative histogram of the contextual amplitude (D. Median ± SE ΔZ-score*s: activated=0.169±0.005, narrow=0.176±0.002, broad=0.185±0.001), duration (E. Mean ± SE last bin ms: activated=252.42±5.76, narrow=246.01±2.03, broad=252.90±1.83) and contextual coverage (F. Median ± SE, percent significant: activated=2.00±0.62, narrow=3.91±1.10, broad=2.14±0.59) for the classified cell types. The insets show the median for amplitude and coverage, and mean for duration, whit the 100-jackknife confidence interval (ns: non-significant, *: p<0.05, ***: p<0.001. Kruskal-Wallis with post hoc Dunn test).

Both optotagged and narrow spiking neurons showed contextual effects of significantly reduced amplitude relative to putative pyramidal cells (Figure 14 D). However, a difference in duration was only significant for inhibitory neurons identified by waveform classification (Figure 14 E). This discrepancy might be a consequence of viral manipulation, or the relatively small number of optotagged neurons. Despite having smaller contextual effects, putative inhibitory interneurons showed contextual effects more often. That is, they showed modulation over a greater proportion of the context space (narrow=10.7+-1.2, broad=8.9+-0.5, activated=5.2+-1.1 percent significant, median and 100-jackknife confidence interval. Kruskal Wallis p=0.02, Dunn post hoc: narrow vs broad p=0.02, all other comparisons non-significant. Figure 14 F).

Pupil effects

Internal states like attention, arousal and task engagement have been associated with changes in the representation of sounds by the auditory cortex (S. V. David et al., 2012; Saderi et al., 2021), through mechanism like increase in firing rate and desynchronization of neuronal activity (Schwartz et al., 2020). We explored the relation of pupil dilation, a correlate of arousal, with the magnitude of contextual effects.

To quantify the effect of pupil on firing rate and context effects, we calculated pupil modulation index (MI) for both probe responses, independent of context (MI-fr), and context effect amplitude (MI-ce) in non-overlapping, 250ms intervals (A:0-250, B: 250-500, C: 500-750, D: 750:1000 ms following probe onset Figure 15 D). We observed a significant positive MI-fr during large pupil (Mean MI-fr: A=0.057, B=0.065, C=0.069, D= 0.079. p<0.001 for all time intervals, one sample T-test). This change recapitulates previous findings of higher firing rate in states of high arousal

||REFS||. The dependence of context effects on pupil size was more complex. Context effects on large pupil trials were larger during interval A, early after the probe onset (MI-ce>0). However, at later time points large pupil was associated with a reduction in context effect amplitude (MI-ce<0. Mean MI-ce: A=0.016, p<0.001; B=0.001, p=0.7, C=-0.019, p<0.001, D=-0.042, p<0.001. one sample T-test).

We also compared MI-fr and MI-ce across the dataset to determine if they were correlated within contextual instances. These difference effects of pupil size were significantly correlated during early intervals but not late intervals (A: r=0.17, p<0.001; B: r=0.11, p<0.001; C: r=-0.01, p=0.5; D: r=-0.03, p=0.3. Pearson's correlation and Wald test for nonzero slope). Activity during interval A may partially reflect an offset response to the context stimulus. The observation that MI-ce during this interval is correlated with MI-fr is consistent with a direct modulation of sound evoked activity by pupil size. The decrease in context effects for large pupil during later intervals is not correlated with MI-fr, suggesting that this change arises from a different mechanism, including an interaction of pupil effects and context information in the local population (see below).

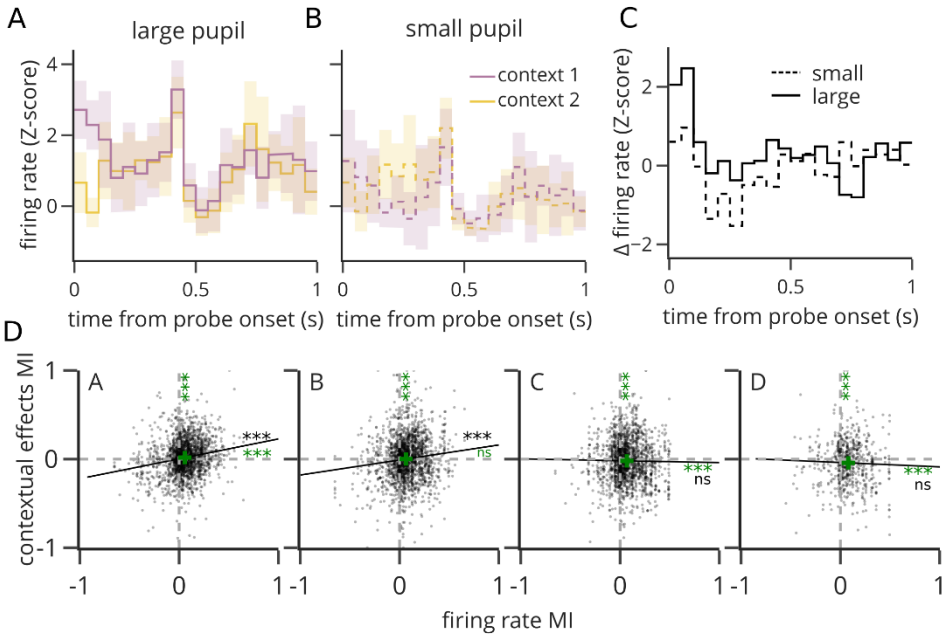

Figure 15. Large pupil is correlated with changes in contextual effects.
A,B. Example PSTH response of one neuron (line: mean, shade: SEM) to a probe sound following two contexts (color) for trials with large (d, solid lines) and small (e. dashed lines) pupil. When pupil is small,

an overall reduction in firing rate pushes both responses towards zero and reduces the amplitude of early contextual effects. **C.** Context-dependent difference (purple – yellow) for large (solid line) and small (dashed lines) pupil trials of the same neuron show in panels A,B. **D.** Comparison between pupil Modulation Index of firing rate (MI-fr) and context effects (MI-ce) at juxtaposed 250ms time intervals A to D after probe onset (the number of dots displayed has been decimated for clarity). Vertical and horizontal dashed gray lines indicate MI=0 (no pupil effect). Green crosses indicate means on x and y. Significant difference from zero on x and y is indicated with green asterisks on the top and right sides ($p<0.001$. ***: $p<0.001$, ns: non-significant. One sample T test). Black lines indicate linear regression with associated significance indicated (Pearson's r. ***: $p<0.01$, ns: non-significant. Wald test for nonzero slope).

Encoding model analysis indicates a role of population activity in representing context

Next, we used an encoding model approach to evaluate mechanisms underlying the observed contextual modulation. We hypothesized several possible mechanisms that could contribute to the context effects: (i) long-latency receptive field properties, (ii) feed-forward adaptation to sensory inputs, (iii) and modulation by the local neural population. To evaluate the role of these mechanisms, we fit a set of generalized linear models (S. V. David, 2018; Thorson et al., 2015) that successively incorporated terms to account for different sources of modulation (Figure 16 A). The STRF model described the activity of a neuron based only on the sound spectrogram. This provided a baseline, reflecting a standard model of sound encoding in AC (Atiani et al., 2014; deCharms et al., 1998). To account for possible adaptation effects, the Self model included an additional input based on the past spiking activity of the neuron being modeled. To account for local population effects, the Pop model incorporated past activity of the simultaneously recorded neurons. Finally, the Full model incorporated both the neuron's own activity as well as the local population activity.

This set of models can be described as an STRF plus optional inputs for self- and population history. To balance parameter count between models, all models contained parameters for the addition inputs, but the inputs were temporally scrambled when the term was not included in that model. Because all models had the same number of free parameters, differences in their performance were a direct consequence of value of the predictors, and not differences in estimation noise.

Models were fit using a separate set of natural sounds that were presented to the same neurons but not used to measure context effects. We then used the models to predict the time-varying spike rate response to the context-probe stimuli. Overall performance was calculated as the Pearson's correlation between real and predicted activity. This performance measurement was agnostic to context and probe classification or transitions. Models incorporating past neuronal activity as predictors performed significantly better (Figure 16 C, Full>Pop>Self>STRF).

To measure our models' ability to capture contextual effects, we compared the amplitude of contextual modulation to between real and predicted responses. Due to the deterministic natures of our models, we could not measure contextual modulation for the predictions using the same T-Score cluster mass method as for the actual data. Instead, we simply measured the amplitude of the difference in predicted probe response between context conditions. To discriminate the temporal profile of contextual effects captured by the models, we computed the amplitude of context differences separately for non-overlapping 250ms intervals (intervals A-D, Figure 16 B). Consistent with the overall measures of prediction accuracy, models with access to information about history of neural activity better predicted the amplitude of contextual effects at all time intervals: The baseline STRF model could capture some contextual modulation soon after probe onset (interval A), but it failed to capture later effects. In contrast, the full model captured these later contextual effects, with peak performance during interval B, between 250 and 500ms (Figure 16 D), but with significant improvement also during later intervals C and D (Figure 16 E).

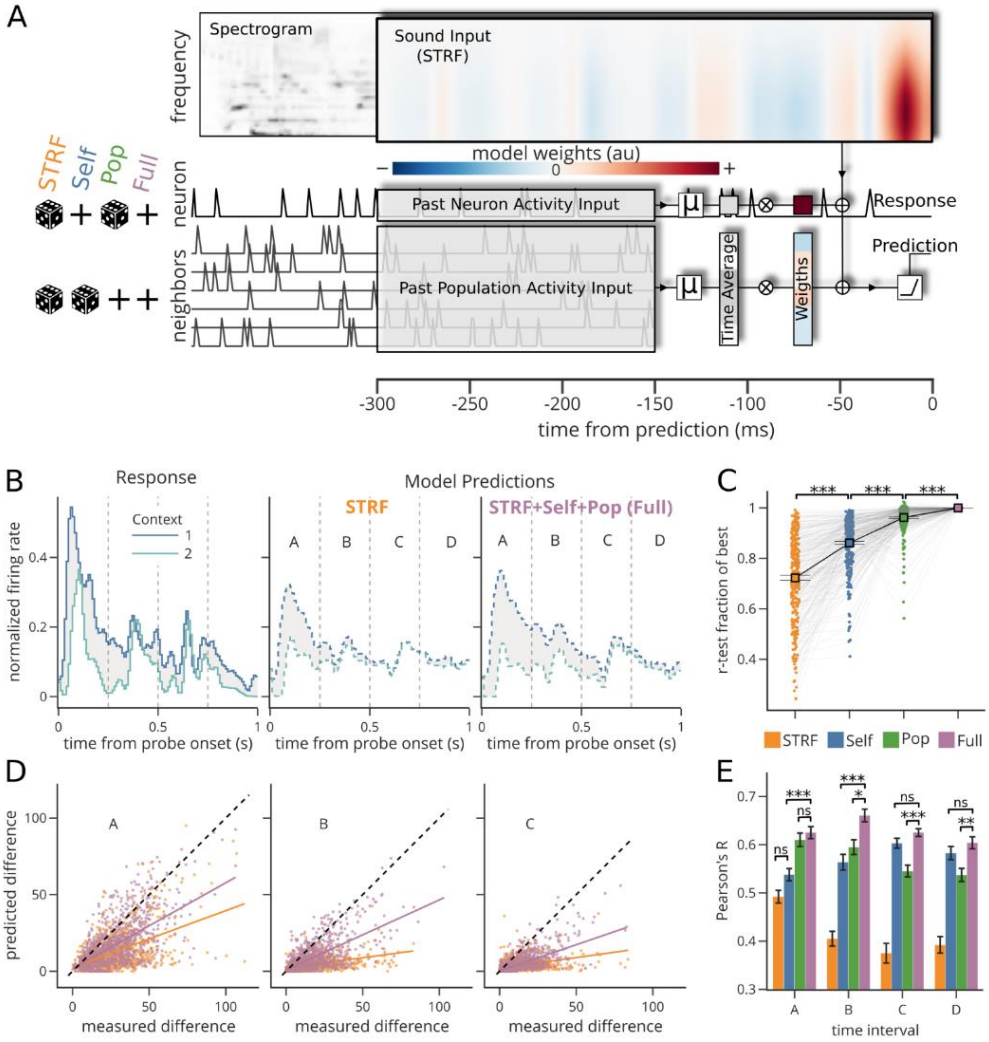

Figure 16. Contextual effects are supported by local population.
A. Architecture of encoding models predicting neuronal responses as a function of sound and the past activity of the neuron (self) and its neighbors (pop). Sound information (spectrogram) was weighted with a spectrotemporal receptive field (STRF) with 18 spectral channels over a period of 300 ms prior prediction. Past activity of the predicted neuron and its neighbors was time averaged over a window extending from 150 ms to 300 ms prior prediction and multiplied with neuron specific weights (colored vectors). The weighted sound and past neuronal activity were summed and passed through a rectifying linear unit (ReLU) to generate the prediction. To test the predictive value of past neural activity, these inputs were time-shuffled (dice) or not (plus sign), defining four different models (left side glyphs. STRF: both shuffled, Self: population shuffled, Pop: neuron shuffled, Full: no shuffling). **B.** Example neuron response (left) and predictions for the STRF (center) and Full (right) models to a single probe after two contexts (blue, teal). The contextual effect (gray area) was calculated for 250ms intervals (A to D). **C.** Quantification of different models' prediction accuracy (cross validated Pearson's r) as a fraction of the Full model accuracy. Colored circles connected by gray lines represent models fit to individual neurons (n=275. Black squares and error bars are the mean and SEM with a corresponding trend line (STRF vs Self: $p<0.001$, Self vs Pop: $p<0.001$, Pop vs Full $p<0.001$. Wilcoxon signed-rank test, Bonferroni corrected). **D.** Comparison of context-depended difference measured and predicted by the STRF (orange) and Full (purple) models, at time intervals A (left) B (center) and C (right). Data wase pooled across all significant instances for all fitted neurons(n=15176). For display clarity dots were decimated by taking a random 1000 points subset. The linear regression and Pearson's correlation coefficients were calculated over all the data (colored lines). **E.** Pearson's correlation coefficient as calculated in d**.** for all

models and time intervals (A-D). Error bars are the standard deviation calculated from the 200-fold jackknifed distribution of the Pearson's r. Model were compared within each time interval. Comparisons not showing in the figure were significant with p<0.001 except A-Self vs A-Pop p<0.01, B-Self vs B-Pop p=0.8, C-Self vs C-Pop p<0.01 and D-Self vs D-Pop p=0.1 (ns: non-significance, *: p<0.05, **:p<0.01, ***:p<0.00. Student's T-test with Bonferroni correction)

The Self and Pop models, which incorporated only one history term, also performed better than the baseline STRF at most time intervals. Moreover, the Self and Pop models approached the Full model performance during different intervals. The Pop model explained most of the performance gain of the full model early (Pop vs Full, interval A, non-significant difference), but the Self model matched the Full model during later intervals (Self vs Full, interval C and D, non-significant differences). This dynamic suggests distinct roles for population activity and the neuron's own history (adaptation, plasticity), with population dynamics carrying most contextual information early on, and later being superseded by the neuron's internal state.

Discussion

A sparse representation of natural auditory context

We observed integration of preceding context stimuli in ongoing sound representations in auditory cortex, lasting hundreds of milliseconds. Neurons showing this integration often only did so under the right stimulus conditions, i.e., differences occurred only for very specific combinations of two context sounds preceding a probe. However, across a population of simultaneously recorded neurons, different neurons showed distinct context sensitivity. The diversity of effects across neurons supports a sparse code for contextual integration at the population level, where the population activity is modulated by more stimulus combinations than any of its constitutive neurons. The sparse representation of significant contextual effects is reminiscent of sparse codes observed for the receptive fields of neurons in the visual, somatosensory and auditory cortices (Lyall et al., 2021; Olshausen & Field, 1996; Zhang et al., 2019). Sparse codes are an efficient representation strategy, which yields independent representations that can be read out during decoding and are redundant and robust (Beyeler et al., 2019).

Effects of context transition types

These contextual effects depend on the spectrotemporal differences between contexts and probe. The amplitude of effects was greater when one of the contexts was silence, and smaller when it was the same sound as the probe. This pattern is reminiscent of the response to oddball or standard sounds traditionally used to study stimulus-specific adapation (SSA). We argue that the contextual effects we observe here are a superset of SSA, since the natural sounds used in our experiments have greater complexity than the parametrically controlled oddball experiments (Natan et al., 2015; Yarden et al., 2022). It follows that the mechanisms underlying SSA, i.e., short term synaptic plasticity, distinct synaptic input, top-down control (Malmierca et al., 2015) are likely to participate in generating contextual effects.

Context effects are larger in non-primary fields of auditory cortex

Secondary auditory regions showed greater contextual effects, in amplitude and to a lesser degree in duration. This result supports the view of a hierarchically organized auditory cortex. More abstract representations emerge at secondary regions following longer and more complex integration (Atiani et al., 2014; Bizley et al., 2005; Norman-Haignere et al., 2022), which is consistent with stronger contextual effects. The regional differences in contextual effects were more prominent in their amplitude rather than duration. We speculate that this difference is due to an increase in the weighting of past vs ongoing stimuli in PEG, rather than a simple increase in the integration window. It is worth noting that PEG is a relatively early secondary region, which shares some tuning features with A1 (Atiani et al., 2014). Our findings suggest that one of these gradual shifts is the stronger weighting of auditory context in PEG representations.

Inhibitory interneurons

Inhibitory interneurons tended to show greater context space coverage, but context effects with smaller amplitude than pyramidal neurons. This seemingly paradoxical behavior can be reconciled if we consider that inhibitory interneurons pool the activity of the local neighborhood, a

phenomenon implicated in normalization of local activity levels (Carandini & Heeger, 2012). We hypothesize that pooling increases the size and complexity of inhibitory interneuron receptive fields (A. K. Moore & Wehr, 2013), thus making them respond to more sounds and contextual interactions. Meanwhile, the same pooling might keep the inhibitory interneuron in a homogeneous state of adaptation, less likely to respond strongly to changes in sound and produce large contextual effects. However, the distinct connectivity and response to thalamic input, with quickly depressing and facilitating PV and SOM interneurons respectively (Tan et al., 2008), might play a role in their expression of contextual effects.

Pupil and arousal promote edge detection but not integration

An increase in pupil-indexed arousal modulates contextual effects in AC bimodally, increasing their amplitude soon after the probe onset (0-250m), and decreasing them later (500-1000ms). We hypothesize that a mode of heightened attention promotes the detection of novelty, which explains the early (probe onset) increase in amplitude. However, to detect novelty, the brain must quickly adapt to the recognized sound statistics to prepare for new deviations. This quick return to a baseline translates into reduced contextual effects at later timepoints. A prediction for this hypothesis is an increase in the response to deviant sounds on the classic SSA-oddball experiment during large pupil trials. This experiment, to our knowledge, has not yet been performed.

Preceding neuronal population activity explains contextual effects

Our models demonstrate that recent past neuronal activity, both from the predicted neurons and its neighboring neurons, is implicated in generating contextual effects. These modes offer an algorithmic explanation rather than a physiological implementation. However, we can draw connections between these two. The effect of the past activity of the predicted neuron relates to its short-term plasticity, where positively weighted past activity relates to potentiation and negative to depression, and therefore to the multiple physiological mechanisms, like synaptic plasticity,

working at in this temporal regime (150–300 ms). The predictive value of past population activity early during probe response, can be related to feedforward computation (Wehr & Zador, 2003), whereas the latter effects are likely to be consequence local recurrent computation (Carandini & Heeger, 2012; Oldenburg et al., 2022), and longer loops through the midbrain (Malmierca et al., 2015) or other cortical regions.

Methods

Animal preparation

Adult male ferrets over 6 months old were surgically implanted with a head post to stabilize the head and enable multiple small craniotomies for acute electrophysiological recordings. Anesthesia was induced with ketamine (35mg/Kg) and xylazine (5mg/kg) and maintained with isoflurane (0.5-2%) during the surgical procedure. The skin and muscle on top of the cranium were removed and the surface of the skull was cleaned. Ten to twelve small surgical screws were placed on the edge of the exposed skull as anchor points. The surface of the skull was chemically etched (Optibond Universal. Kerr) and a thin layer of UV-cured dental cement (Charisma Classic. Kultzer) was applied over the exposed surface. Two stainless steel head posts were aligned along the midline and embedded with additional cement. Finally, cement was used to build a rim extending out from the edges of the implant. The rim served the dual purpose of holding bandages over the implant margin wounds and creating wells to hold saline over the recording sites. Once the implant was finished, excess skin around it was removed, the wound around the implant was closed with sutures and the animal was bandaged. Antibiotics and analgesics were administered as part of the post-op recovery.

After >2 weeks following surgery the animals were acclimated to a head-fixed posture, during intervals starting at 5 minutes and increased 5 to 10 minutes every day. Food and liquid rewards were given during these acclimation sessions to help the animals relax under restraint. Animals

were considered ready for recording when they could be restrained for more than 3 hours without signs of distress (e.g., the animals being relaxed enough to fall asleep).

Sound presentation

Acoustic stimuli were either synthesized or drawn from a library of pre-recorded samples and presented using custom Matlab software. Digitized signals were converted to analog (National Instruments) and amplified (Crown). They were presented to head-fixed animals in a sound attenuating chamber (Gretch-Ken or Acoustic Systems), using calibrated free-field speakers (Manger) positioned at 30-deg contralateral azimuth, 0-deg elevation, and 80 cm distant from the animal.

Auditory stimuli used for measuring sensory context effects were sequences of 1-sec natural sounds, which we refer to as context-probe pairs. Sequences were constructed so that each probe sound was preceded by several different context sounds. To maximize efficiency of context-probe sampling, we generated sequences of N different sounds, such that any sound acted as the probe following a preceding context, or as the context for the following sound (Asari & Zador, 2009). Each sound was also played at the beginning of the sequence, therefore acting as a probe following a silent context. Sounds were also repeated so that a probe could also provide its own context.

For N different sounds, full sampling of context-probe combinations was achieved with N sequences of N+1 sounds (Figure 11 A)(Asari & Zador, 2009). Finding sound sequences fulfilling these conditions poses a mathematical problem known as “exact coverage”, which we solved using the dancing links algorithm (Knuth, 2000). We created sequences from N=4 or 10 1-second sound samples. In each experiment, sounds were drawn from a set of 16 natural sounds, based on their ability to drive neuronal activity in the recording site. This 16-sound set was chosen from a large library, selected for their ability to drive activity across many neurons in previous recordings in the laboratory. It contained music, speech, ferret vocalizations and environmental noise such as gravel and brushes. We defined three broad categories of context based on their

relationship to the probe: silence, a different sound than the probe, and the same sound as the probe (Figure 11 C).

Neurophysiological recording

The putative location of A1 and PEG was determined during the implantation surgery based on external landmarks: the posterior and medial edges of A1 falling, respectively, 13 mm anterior to the occipital crest and 8 mm lateral to the center line, and PEG immediately antero-lateral to A1 (Bizley et al., 2005). To functionally confirm recording locations, we opened small craniotomies of ~1mm diameter and performed preliminary mapping with tungsten electrodes (FH-Co. Electrodes, AM Systems Amp, MANTA software (Englitz et al., 2013)). We measured the tuning of the recording regions using rapid sequences of 100ms pure tones and used tonotopy to identify cortical fields (Bizley et al., 2005). We specifically looked for the frequency tuning inversion: high-low-high moving in an antero-lateral direction, which marks the boundary between primary (A1) and secondary (PEG) fields. At tonotopically mapped sites, we performed acute recordings with 64-channel integrated UCLA probes (Du et al., 2011), digital head-stages (RHD 128-Channel, Intan technologies) and OpenEphys data acquisition boxes and software (Siegle et al., 2017).

Raw voltage traces were processed with Kilosort 2 (Stringer, Pachitariu, Steinmetz, Reddy, et al., 2019), clustering and assigning spikes to putative single neurons. The clusters were manually curated with Phy (Rossant et al., 2016). Units were only kept for analysis if they maintained isolation and a stable firing rate over the course of the experiment. Unit isolation was quantified as the percent overlap of the spike waveform distribution with neighboring units and baseline activity. Isolation > 95% was considered a single unit and kept for analysis. We further filtered neurons based on the reliability of their responses, requiring a Pearson's correlation > 0.1 between PSTH responses to the same stimuli (10 repetitions, 20 Hz sampling) drawn from random halves of repeated trials.

Evaluating significance of sensory context effects

To measure effects of sensory context on sound-evoked activity, spike times for each unit were binned at 20 Hz. Activity was normalized as a z-score based on mean and standard deviation of single-trial spike rate across the entire duration of the recording (spontaneous activity and during sound presentation). We define a contextual instance as a probe sound preceded by a pair of context sounds. Experiments using N=4 distinct sound samples produced 50 distinct contextual instances, and experiments using N=10 sounds produced 550. For each contextual instance, we computed the difference in the response to the probe between the two contexts. To track contextual effects over time, we calculated this difference at every 50-ms time bin (Δ Z-score).

To evaluate the significance of differences in spike rate, traditional analysis might use a T or U test for the response difference at every time bin. However, this approach leads to the problem of multiple comparisons or reduced sensitivity if using the Bonferroni corrections across many time points in the probe response. To maximize statistical power, instead, we used a cluster mass quantification of significance (Maris & Oostenveld, 2007), which corrects for multiple comparisons in the time domain, without sacrificing sensitivity.

For each contextual instance, we first calculated the T-score during each time bin of the probe response between each context. We then found groups of one or more contiguous time bins with significant T-scores of the same sign ($p<0.05$). Each of these groups defined a cluster with an associated score computed as the sum of the T-score for all time bins in the cluster. Each cluster, finally, was assigned a p-value calculated by comparing the cluster score to its null distribution. This null distribution was obtained by calculating the maximum cluster statistic value (following the same procedure as above) for 11000 random shuffles of the context identity (Figure 11 F, J). We calculated this cluster-mass T-score and p-value for all contextual instances for a given neuron. We used the Bonferroni method to correct for multiple comparisons across contextual instances with a family error of alpha=0.05.

Amplitude and duration of context effects

The temporal profiles of the context dependent differences in firing rate were diverse and irregular; therefore, we avoided describing them with monotonic distributions, e.g., exponential decay. Instead, we quantified the amplitude of contextual differences as the sum of the absolute difference (|Δ Z-score|) across significant time bins and their duration as the time of the last significant time bin (Figure 11 G, K).

Context type and region effect

We performed categorical multivariate linear regression to quantify the dependence of the amplitude and duration of contextual modulation on context type (Silence, Different, Same) and cortical region (A1, PEG). The amplitude and duration metrics for each contextual instance were normalized by dividing by the grand mean amplitude or duration across all neurons and contextual instances, thus scaling them to the percent change relative to average. Most contextual instances were comprised of two Different contexts. Since Different was the most common context, it was assigned as the base, dummy variable. Categorical inputs for Silence and Same were then set to a value of 1 for instances that included one or both of these categories. Significance of the regressed coefficients was quantified with a T-test over the residuals of the regression. Linear regression, using Ordinary Least Square minimization, and significance statistics were calculated using the python package Statsmodels (Seabold & Perktold, 2010).

Because the distributions of contextual amplitude and durations were strongly skewed and non-Gaussian, we validated the significance of context type and region differences with a non-parametric ANOVA (Kruskal-Wallis) and Dunn post hoc tests.

Sparse population coding analysis

For every neuron, the significant contextual modulation at every contextual instance yielded a coverage of the contextual modulation space (all probes by all context-pairs, Figure 13 A). In

each site, the best neuron had the greatest contextual coverage, i.e., had the greatest number of contextual instances with significant contextual modulation.

To describe the optimal contextual coverage of a site we took the union of the contextual coverage of the individual constitutive neurons (Figure 13 B top). For every contextual instance, we took the value with the highest amplitude amongst all neurons. For this union we also considered neurons as another source of multiple comparisons and corrected for it alongside the prior correction for number of contextual instances.

To describe contextual modulation of population activity in a low dimensional space, we used Principal Components Analysis (PCA, Figure 13 B bottom). We fitted the PCA transformation matrix on trial-averaged responses to n-sound sequences for all neurons in a recording site. Thus, we included all the information available about sensory responses, while eliminating trial to trial variations. We used the transformation matrix to project single-trial responses onto the first principal component. These projections were then used to calculate contextual modulation following the same procedure as for single neurons.

Viral injection

For one animal, we injected into the cortex an adeno-associated virus serotype 2 (AAV2) containing channelrhodopsin 2 (ChR2), and mCherry under the inhibitory interneuron specific promote mDlx (Dimidschstein et al., 2016). Two craniotomies were drilled and adequate injection sites spanning A1. The injection locations were validated as for electrophysiology. The injections were performed under Ketamine-Xylazine anesthesia and vitals were tracked through the procedure. The animal head was fixed using the previously implanted headcap. A glass injection needle with a beveled tip of ~30µm diameter, was coupled with flexible tubing to a 100µl syringe (Hamilton, 7656-01), controlled with an automated injector (New era pump systems, NE-1000). The syringe was preloaded with mineral oil (Sigma-Aldrich), which was used to prime the whole hydraulic system. 10µl of virus were back loaded, and ~5µl were injected in each craniotomy. The injections were performed at a depth of ~1.5mm, roughly in the middle of the cortical depth. To

improve the coverage of the viral injection, we used a convection enhance delivery strategy (Weiss et al., 2020), where the delivery rate started from 0.5µl/min and was incremented by 0.5µl/min every 3 minutes until the desired volume was injected. The incubation period between injections and photo stimulation was 2 weeks.

Optotagging

To photo stimulate and record from neurons simultaneously we attached an optic fiber (24 mm, 0.66 NA, 400µm inner diameter, 430µm cladding, 1.25 zirconia ferrule. Doric lenses MFC_400/430-0.66_24mm_ZF1.25(G)_FLT) to UCLA 64 channel probe using nonconductive, encapsulating epoxy (Resinlab EP965). The optic fiber laid parallel and in contact with the probe shank, leaving ~1.5mm of clearance between the fiber face and the electrodes. This clearance was enough so the array could be introduced into the cortex, and the optic fiber would lay on top of, or close to the dura. the ferrule was connected to a laser (Ikecool, IKE-473-100-OP) with an optic patch cord (2.5m, 0.57na, 400 µm inner diameter, 430 µm cladding. Doric lenses MFP_400/430/3000-0.57_2.5mm_FC-ZF1.25). Laser power delivered close to the dura was calibrated between 200-250 mW/mm2.

The photo-stimulation consisted of 40 trials of a single 20ms flash delivered during silence. The inter trial interval was 1s, and the flash trials were randomly interspersed with control trials with no light. The light stimulation generated a significant photoelectric artifact consisting of high amplitude and low duration on and offset transients, and a sustained lower amplitude noise. The transients were eliminated by removing the 2ms right after laser on and off and interpolating to fill missing values. The lower amplitude ongoing noise was reduced by subtracting the common average of laser trials. Preprocessed data was then spike sorted as before, and the remaining artifacts appearing as spike clusters or outlier spikes on good cluster were discarded. Neurons were considered opto-tagged if they responded within 5ms to the light onset, with a train of action potentials reliable across trials (Figure 14a).

Spike wave form analysis

Neurons were classified based on their average spike waveform width, which was calculated as the time between the depolarization valley and the hyperpolarization peak (Trainito et al., 2019). The spike width was calculated for all neurons with amenable waveforms (inverted mostly positive waveforms with multiple inflections, associated with axonal spikes (Sibille et al., 2022), were difficult to interpret and excluded). The distribution of spike widths followed a clear bimodal distribution. The width threshold was defined as the valley in the kernel density estimation, and a safety range of 0.1 ms around the threshold was kept as unclassified.

Pupillometry

To obtain a measure of global changes in arousal, pupil was recorded during experiments with a video camera (Adafruit TTL Serial Camera 397, M12 Lenses PT-2514BMP 25.0 mm) placed 10 cm from the eye. An infrared light was used to improve the contrast of the image. The pupil was kept partially contracted with an ambient light set to ~1500 lux at the eye being recorded, this increased the dynamic range of pupil size. The pupil size was measured offline using software detailed in (Schwartz et al., 2020).

Trials were classified by the median pupil size yielding a balanced number of large and small trials. This classification was performed independently for each contextual instance, considering the mean pupil size across the time interval containing both context and probe.

The context-independent pupil-dependent firing rate (mean Z-score) was calculated for all combinations of neurons and probes, averaging across contexts (Figure 15 A), and over the 1s probe duration. Nonresponsive neuron-probes (mean Z-Score<0.1) were filtered out for further analysis.

The pupil-dependent contextual modulation (mean Δ Z-score) was calculated for significant contextual instances identified previously using the cluster mass analysis (see above). The magnitude of the pupil dependent effect was computed by averaging the Z-scored contextual modulation across the entire probe response. This metric was also calculated for non-overlapping

250ms time intervals (A:0-250, B:250-500, C:500-750, D:750-1000). Values with low firing rate (mean Z-score<0.1) and small contextual modulations (mean Δ Z-score<0.3) were filtered out. Prior to computing pupil effects, the sign of contextual modulation was flipped to be positive, so that the corresponding pupil-dependent contextual modulation was also positive on average.

A pupil modulation index (MI) was calculated for firing rates and contextual modulation. MI was defined as (large – small) / (|large| + |small|), where |·| denotes the absolute value of the relevant statistic. This rectification accounted for instances where the value was close to zero and fell to a negative value for one pupil condition. MI close to zero indicates no pupil effects, while negative and positive values indicate an increase for small and large pupil respectively.

Model architecture

We trained encoding models to predict the activity of a neuron as a function of sound stimuli, its own past activity and that of other neurons in the population. All models followed the same Generalized linear model architecture:

Input sound was transformed into a log-spaced, 18-channel spectrogram (approximately 1/3 octave per channel) with amplitude log compression emulating cochlear dynamics (Pennington & David, 2022). Stimulus and spike signals were binned at 100 Hz. Spike data was averaged across trials and normalized to the peak value for each neuron. We used a standard linear-nonlinear spectro-temporal receptive field (STRF) as a baseline model of sensory encoding. The STRF spanned the 18 spectral channels over a window 300 ms (30 10-ms bins) before the neuron response. A predicted response was computed by treating the STRF as a filter, convolving with the stimulus spectrogram in time and summing across spectral channels (Figure 16 A, top STRF).

To model the effect of the past neuronal activity, the response of all recorded neurons (including the neuron being predicted) was read over a time window spanning150-300 ms before the neural response. Neuronal activity was averaged over this time window, and the weighted sum of these averages was added to the STRF output (Figure 16 A, past neuron, and population activity). Using the neuronal activity time average instead of every time point reduced the number

of parameters, making the model more interpretable, where every parameter is an average synaptic strength in the population.

Traditional STRFs of 150 ms can best capture the auditory driven responses in A1 and PEG (Atiani et al., 2014), which happen in that time regime. Therefore, we used the population filters that were offset 150 ms into the past to avoid capturing correlated sound evoked activity in the other neurons, and rather focus on the effect of recent population activity, i.e., a proxy for network activity. To disambiguate changes in model performance due to a longer temporal window extended by the population filter, we also extended the STRF to 300 ms into the past, so its first half overlapped with the population filters. Ideally these “far past” sounds should carry little information about the sound evoked response, therefore, these weights tend to zero, and the nonzero weight of the STRF remain at short time lags (< 150 ms). Finally, the summed output of the STRF and the neural filters was then passed through a rectified linear unit (ReLU) to account for spike threshold (Thorson et al., 2015).

We defined 4 different models based on this same architecture, by temporally scrambling different parts of their input: 1. a base STRF, achieved by scrambling both the self and population response 2. a “Self” model, where only the other neurons response was scrambled 3. a “Population” model (pop), where the self-response was scrambled 4. a Full model, with no scrambling (Figure 16 A, left sigils).

This scrambling effectively removes the predictive value of the scrambled data, while keeping the total number of parameters unchanged. Thus, since all models have the same number of parameters, a direct comparison of their performance is valid (Figure 16 C). Temporal scrambling was done by independently shifting neuronal responses by random tens of seconds. We chose to shift instead of shuffling the data to keep the short-term temporal structure of the data and prevent the models from fitting to the grand average of the scrambled predictors.

All models were implemented using NEMS (Pennington & David, 2022), a flexible and readily available software developed in the lab. Optimization was performed using the ADAM gradient descent algorithm (Kingma & Ba, 2017).

Model Performance quantification

Model performance was quantified as the prediction correlation, computed as Pearson's R across a predicted response assembled from N cross-validated model estimates (N=4 or 10, the number of different stimulus sequences).

To further quantify the ability of the different models to account for long-lasting contextual effects, we compared the context driven difference for neuronal responses and model predictions in each contextual instance. For the predictions, context modulation was computed similarly to the sum of absolute Δ Z-score calculated from the actual response. However, there are two main differences. First, we used 0 to 1 normalization, instead of Z-scores, as models performed better this way. Second, there was no associated significance test (e.g., cluster mass analysis), due to the deterministic nature of our models.

We evaluated the accuracy with which models predicted contextual effects by calculating the Pearson correlation between the amplitude of contextual effects in the real data and those predicted by the model. This correlation was also calculated separately for 4 non- overlapping 250ms time intervals spanning the probe response, named A to D (Figure 6 d, e). This analysis was performed only on contextual instances that were significant according to the cluster mass test performed on the original data.

Chapter 4. Conclusions and Future Directions

Coordinated cortical integration windows

Previous work has demonstrated the existence of long-lasting temporal integration in the auditory cortex from the perspective of individual neurons (Asari & Zador, 2009) or local field

potentials (Norman-Haignere et al., 2022). However, how this integration contributes to the representation of natural sounds by populations of neurons has remained unexplored.

Here we demonstrate that adaptation, a mechanism tied to temporal integration, happens independently for different spectral inputs to AC neurons. This spectrally tuned adaptation contributes to a diversity of integration properties across neurons, which is necessary to capture the rich spectrotemporal patterns present in natural sounds. Furthermore, we studied how adaptation of individual neurons interacts though local connectivity, giving rise to population dynamics and diverse integration windows and sound representations across sensory contexts.

Our results recapitulate prior quantifications of temporal integration windows. However, by looking across a population of neurons and a diversity of natural stimuli, our findings show complex and variable integration. During natural sound processing, the representations of context that emerge from this integration are distributed through the population in a sparse code. Finally, we also show that these integration properties depend on cortical region, cell type, and the spectrotemporal characteristics of the sounds being integrated.

A loose definition of context

We have treated context as the recent, < 1 sec acoustic history in which sound has to be interpreted. This definition, however, is just a sliver of a much broader setting in which hearing occurs. A holistic view of context needs to include sound history at longer time scales, of tens of seconds and beyond (Bianco et al., 2020; K. Lu et al., 2018a). It also must consider the context of background or competing sounds that occur simultaneously (Micheyl et al., 2007; R. C. Moore et al., 2013). More broadly, sound acquires different meanings based on a multisensory context (Choi et al., 2018), which defines space, our position inside it, and the embodied percept of sources. Hearing the roar of a tiger coming from a speaker at the comfort of our desk is very different from an invisible source in a dense jungle.

Any meaningful distinction between these multiple aspects of context should be found in the mechanisms and physiology underlying them. For example, it is likely that representations of the recent temporal context and the context of background noise are supported by the same mechanisms of adaptation, circuit connectivity, and hidden population states (Buonomano & Maass, 2009). Therefore, these two contexts are probably strongly coupled. We ought to study both processes together. Conversely, the recent temporal context, likely encoded in the local hidden state of a cortical column, will be loosely tied with an longer-lasting temporal context, likely captured by the recurrent interactions between regions in the auditory pathway (Heilbron & Chait, 2018; Malmierca et al., 2015).

In our current approach, we have focused exclusively on the recent temporal context, however we could extend this approach to capture the context other sound playing simultaneously. This can be achieved by overlapping parts of contexts and probes, instead of simply juxtaposing them.

Considering the hierarchical organization of temporal integration along the auditory pathway, we can investigate the emergence and interactions between time scales by recording the coordinated activity across sequential regions of the auditory pathway. This has been recently enabled by improvements in our recording paradigm using neuropixels (~1cm), with a length that permits simultaneous recordings from cortex and subcortical regions: A1 and MGB with precise positioning.

New protocols for chronic recording also permit studying neural activity in freely moving animals. Free moving experimental approach enables questioning multisensory integration (Choi et al., 2018), and particularly the context of behavior and motor feedback, which plays a critical role in attenuating the response to self-generated sound (Schneider et al., 2021). Free moving experiments present a tradeoff between the richness and naturality of behavior, and the difficulty of analyzing an increased number of uncontrolled parameters. This increased experimental complexity can be addressed with novel artificial intelligence (AI) tools, which permit the unbiased quantification of uncontrolled variables like of motion and behavior (Mathis et al., 2018), which

can be readily related to neuronal activity (Syeda et al., 2022). Automated video annotation is just one of the recent developments in AI that will contribute to the analysis of data in ever growing quantities and complexity. However, analysis tools are not the only machine learning development relevant to neuroscience.

Implications of deep learning

We are in a significant historical and technological moment, as artificial intelligence rushes through a renaissance, supported by the development of training algorithms like backpropagation, deep net architectures, and increasingly powerful and specialized hardware, which on consumer GPUs reaches ~$35x10^{12}$ FLOPS. An example of the most recent advances of these renaissances includes large language models and latent diffusion, capable of creating “novel” art based on just a text prompt (Nichol et al., 2022), approaching the flexibility, creativity, and performance of humans in specific tasks.

The success of these models comes in part from the increasing size of datasets used to train them. However, these models cannot be scaled indefinitely: error decreases as a power (or an exponent in the best of cases (Sorscher et al., 2022)) of the model and training set size, i.e., there are diminishing returns, and as model performances increase by small fractions, the model complexity and training dataset become prohibitively big. This brings significant ethical implications. We cannot disregard the carbon footprint, the questionable large scale data mining, and the vast (exploitative) labelling effort (Williams, 2022) required to train these models. The question becomes not if we can, but if we should continue progress on machine learning through scale alone.

Furthermore, the success of these models lies in specific architectures discovered by trial and error. For example, in time series forecasting, recurrent neural networks (RNN) were soon replaced by short and long short term memory (LSTM) networks (Yu et al., 2019), which in time were replaced by transformers (Vaswani et al., 2017). Similarly, in image generation, generative

adversarial networks (GAN) (Goodfellow et al., 2020) were quickly outperformed by latent diffusion (Nichol et al., 2022). To continue this brute force exploration rather than drawing inspiration from the brain is inefficient considering eons of evolution's brute force, parallelized optimization. Departure from biological inspiration is not limited to architecture, but also includes the predominance of continuous rather than spiking networks, and the use of global error assignment – back propagation – instead of local, biologically realizable, Hebbian plasticity.

This departure from neuroscience is recent. Early developments in AI were catalyzed by discoveries in neuroscience which inspired the perceptron, an idealized point neuron (Rosenblatt, 1958), and Hopfield networks reminiscent of the hippocampal architecture (Hopfield, 1982). More recent AI developments have also drawn inspiration from the brain: deep-net architectures, like convolutional neural networks reminiscent of the retinal architecture (LeCun et al., 1989), and transformers, loosely inspired by the concept of attention (Vaswani et al., 2017).

The current neuro-AI schisms need to be reconsidered, and there are already people doing it. New theoretical works look to explain fundamental rules of intelligence in both brains and computers. Some of these ideas posit parsimony and self-consistence as common organizational principles in the brain, and in successful AI architectures. Parsimony as *"… to identify low-dimensional structures in observations of the external world and reorganize them in the most compact and structured way."* and self-consistecy as a *"... model for observations of the external world by minimizing the internal discrepancy between the observed and the regenerated."* (Ma et al., 2022). Interestingly, these principles recapitulate principles of brain activity, like sparseness, and general theories of brain function like the free-energy principle (Friston, 2010) which explains intelligent systems as trying to sustain homeostasis by avoiding unexpected conditions, i.e., minimizing entropy and (sensory) surprise.

Predictive coding

These general intelligence theories have direct connections to experimentally backed brain function theories. Such is the case of predictive coding (PC), which is supported by observations across multiple sensory systems (Carbajal & Malmierca, 2018; Heilbron & Chait, 2018; Rao & Ballard, 1999). In general terms, PC posits top-down activity in the brain corresponds to world predictions, which are continuously trying to match the bottom-up activity elicited by sensing the world, such that when predictions are incorrect, this error is propagated to correct the top-down model of the world.

From the PC perspective, neurons are not representing sound (sensory) features but rather hypotheses and predictions about sound. The temporal integration, and contextual effects we observe, can be interpreted within this framework as deviations from the expected continued sound, and the time it takes the auditory brain to recalibrate its expectations to statistics of the novel sound.

PC states that there is a distinct distribution of neurons representing predictions and errors. In the canonical view, cortical superficial layers sending bottom-up projections contain error neurons, while deeper layer sending top-down projections contain prediction neurons. If contextual effects are an expression of prediction errors, they should be enriched on the surface layers of the cortex (Heilbron & Chait, 2018). We are currently working towards the identification of cortical layers based on current source density and oscillation frequency analysis, which would enable the localization of recorded neurons in the cortical column.

A main criticism of predictive coding has been the lack of evidence for an actual efferent prediction which recapitulates the expected sensory input. If we consider the precise reciprocal wiring required to achieve this, and the clear asymmetry between ascending and descending pathways, the lack of this exact efferent copy seems unavoidable. This vertical asymmetry is present not only in the wiring between regions, but also in the temporal dynamics of transmission:

bottom-up transmission is associated with the fast gamma band, while top-down works on the slower alpha band (van Kerkoerle et al., 2014).

Slower top-down predictions will fail to track faster variation in sensory signals; however, we argue that some of these faster predictions might be locally encoded in the hidden state of the network, better suited for faster computation through local synapses, i.e., the contextual effects we observed. There is a division of temporal labor with local, fast, prediction of small deviation, and a top-down prediction, much slower, but on a much broader space of possibilities. As mentioned above, Recordings of populations of neurons across successive regions in the auditory pathway are necessary to test this hypothetical division, and the contribution of these modes of PC in natural sound representation.

Sparse code

We have described sparseness of contextual effects. However, the specificities of the sounds that elicit these effects, i.e., the dimension(s) of these contextual effects, are to be more exhaustively explored. One possible approach is to identify the receptive fields of changes in sound, or tuning to context, that elicits activity beyond first order (STRF) sound-evoked activity. Finding the specific dimensions is fundamental, as the representations of different sound dimensions will show different degrees of sparseness at different brain regions. For example, the joint dimension of sound phase and frequency, understood as gammatones, are sparsely represented in the auditory nerve (Lewicki, 2002), but frequency by itself, is represented by a tonotopic map of labeled lines, i.e., a local code.

Some attempts at capturing these temporally dynamic receptive fields have made use of deep learning, with model architectures designed such that the weights of the model can be visualized as a classic STRF that is dynamically changing over time, as sound progresses (Keshishian et al., 2020). With these dynamic STRFs (dSTRF), it is possible to determine the fraction of the

receptive field caused by contexts independent of ongoing sound, similar to the idea of a contextual receptive field.

Formally capturing and describing contextual receptive fields is a necessary step to determine sparseness in its strict and classical meaning, i.e., as captured by a generative model with enforced sparseness. Some of the more recent deep sparse modes (Zhang et al., 2019), capable of recapitulating first order STRFs across the auditory pathway, might also recapitulate these contextual receptive fields.

Sound diversity and encoding models

Dynamic STRFs and the deep biological networks in which they are based suffer the same hunger for data as the bigger deep learning models discussed above. Auditory neurophysiological models have the extra limitation that experimental constraints make it difficult to acquire large datasets of neuronal sound responses. There are competing interests for the experimentalists: On one hand, classical approaches required low dimensional stimulus sets, like pure tones or trains of clicks which unequivocally dissect particular response properties. On the other hand, the need to use a sound set that is diverse enough to encompass the fraction of the sound space to which a neuron or population is tuned.

Classically, diversity is achieved with large libraries of natural sounds, likely capturing relevant sound statistics to which neurons evolved in nature. However, this also means playing sounds which do not drive neurons and dismissing well known tuning properties of neurons like their preferred frequency. To efficiently stimulate a neuron, a compromise must be found.

Closed loop in line sound generation

We can leverage deep learning and online closed loop sound generation such that the response of neurons to a sound influence the next stimulus. This establishes an iterative process to explore a sound space, and select sound that better drives a given neuron, or perhaps elicits any arbitrary pattern of activity across neurons in the recorded population. This approach had

some success with on-line closed loop image generation in macaques and mice (Ponce et al., 2019; Walker et al., 2019). However, it is based on GANs, notorious for their instability and tendency to mode collapse, i.e., finding a single solution (image) and sticking with it, thus failing to explore more of the stimulus space. Furthermore, these generator models lacked information of what visual neurons respond on average.

We can leverage historical recordings of sound evoked activity of thousands of neurons across animals, and train newer alternatives to GANs. The expectation is that these models will capture a latent space describing the preferred sound features of neurons at specific auditory regions (Pennington & David, 2022). This assumes common rules of tuning across large pseudo populations of neurons recorded in the same region across multiple recording sites and animals. Moreover, we can make models unconstrained by physiology and interpretability, which will be likely to capture the more elusive nonlinearities that determine complex receptive fields.

Using such a tool we can find very precise sounds that drive specific patterns of activity. This can yield more focused stimulation sound sets, with more repetitions and greater statical power. Furthermore, this can transform the process of understanding neuronal tuning into a more trackable analysis of the spectrotemporal properties of the generated sounds. Finally, it will let us directly listen to what neurons care for.

References

Aertsen, A. M. H. J., & Johannesma, P. I. M. (1981). The Spectro-Temporal Receptive Field—A functional characteristic of auditory neurons. *Biological Cybernetics*, *42*(2), 133–143. Scopus. https://doi.org/10.1007/BF00336731

Ahrens, M. B., Linden, J. F., & Sahani, M. (2008). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *Journal of Neuroscience*, *28*(8), 1929–1942.

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 1043–1047. https://doi.org/10.1126/science.aaz3468

Angeloni, C., & Geffen, M. (2018). Contextual modulation of sound processing in the auditory cortex. *Current Opinion in Neurobiology*, *49*, 8–15. https://doi.org/10.1016/J.CONB.2017.10.012

Asari, H., & Zador, A. M. (2009). Long-Lasting Context Dependence Constrains Neural Encoding Models in Rodent Auditory Cortex. *Journal of Neurophysiology*, *102*(5), 2638–2656. https://doi.org/10.1152/jn.00577.2009

Ashida, G., & Carr, C. E. (2011). Sound localization: Jeffress and beyond. *Current Opinion in Neurobiology*, *21*(5), 745–751. https://doi.org/10.1016/j.conb.2011.05.008

Asokan, M. M., Williamson, R. S., Hancock, K. E., & Polley, D. B. (2021). Inverted central auditory hierarchies for encoding local intervals and global temporal patterns. *Current Biology*, *31*(8), 1762-1770.e4. https://doi.org/10.1016/j.cub.2021.01.076

Atencio, C. A., Sharpee, T. O., & Schreiner, C. E. (2008). Cooperative nonlinearities in auditory cortical neurons. *Neuron*, *58*(6), 956–966.

Atencio, C. A., Sharpee, T. O., & Schreiner, C. E. (2009). Hierarchical computation in the canonical auditory cortical circuit. *Proceedings of the National Academy of Sciences*, *106*(51), 21894–21899. https://doi.org/10.1073/pnas.0908383106

Atencio, C. A., Sharpee, T. O., & Schreiner, C. E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. *Journal of Neurophysiology*, *107*(10), 2594–2603. https://doi.org/10.1152/jn.01025.2011

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent Selectivity for Task-Relevant Stimuli in Higher-Order Auditory Cortex. *Neuron*, *82*(2), 486–499. https://doi.org/10.1016/j.neuron.2014.02.029

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., & Chait, M. (2016). Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences*, *113*(5), E616–E625. https://doi.org/10.1073/pnas.1508523113

Barlow, H. B. (2012). Possible Principles Underlying the Transformations of Sensory Messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 216–234). The MIT Press. https://doi.org/10.7551/mitpress/9780262518420.003.0013

Barral, J., Wang, X.-J., & Reyes, A. D. (2019). Propagation of temporal and rate signals in cultured multilayer networks. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-11851-0

Bendor, D. (2015). *The Role of Inhibition in a Computational Model of an Auditory Cortical Neuron during the Encoding of Temporal Information*.

Beyeler, M., Rounds, E. L., Carlson, K. D., Dutt, N., & Krichmar, J. L. (2019). Neural correlates of sparse coding and dimensionality reduction. *PLOS Computational Biology*, *15*(6), e1006908. https://doi.org/10.1371/journal.pcbi.1006908

Bianco, R., Harrison, P. M., Hu, M., Bolger, C., Picken, S., Pearce, M. T., & Chait, M. (2020). Long-term implicit memory for sequential auditory patterns in humans. *ELife*, *9*, e56073. https://doi.org/10.7554/eLife.56073

Binder, J., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human Temporal Lobe Activation by Speech and Nonspeech Sounds. *Cerebral Cortex*, *10*(5), 512–528. https://doi.org/10.1093/cercor/10.5.512

Bizley, J. K., Nodal, F. R., Nelken, I., & King, A. J. (2005). Functional Organization of Ferret Auditory Cortex. *Cerebral Cortex*, *15*(10), 1637–1653. https://doi.org/10.1093/cercor/bhi042

Brumberg, J. C., Pinto, D. J., & Simons, D. J. (1999). Cortical Columnar Processing in the Rat Whisker-to-Barrel System. *Journal of Neurophysiology*, *82*(4), 1808–1817. https://doi.org/10.1152/jn.1999.82.4.1808

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, *10*(2), Article 2. https://doi.org/10.1038/nrn2558

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. https://doi.org/10.1137/0916069

Calabrese, A., Schumacher, J. W., Schneider, D. M., Paninski, L., & Woolley, S. M. N. (2011). A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. *PloS One*, *6*(1), e16104. https://doi.org/10.1371/journal.pone.0016104

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), Article 1. https://doi.org/10.1038/nrn3136

Carbajal, G. V., & Malmierca, M. S. (2018). The Neuronal Basis of Predictive Coding Along the Auditory Pathway: From the Subcortical Roots to Cortical Deviance Detection. *Trends in Hearing*, *22*, 2331216518784822. https://doi.org/10.1177/2331216518784822

Carlson, N. L., Ming, V. L., & DeWeese, M. R. (2012). Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus. *PLOS Computational Biology*, *8*(7), e1002594. https://doi.org/10.1371/journal.pcbi.1002594

Casado, A., & Brunellière, A. (2016). The influence of sex information into spoken words: A mismatch negativity (MMN) study. *Brain Research*, *1650*, 73–83. https://doi.org/10.1016/j.brainres.2016.08.039

Chance, F. S., Nelson, S. B., & Abbott, L. F. (1998). Synaptic depression and the temporal response characteristics of V1 cells. *Journal of Neuroscience*, *18*, 4785–4799.

Chechik, G., Anderson, M. J., Bar-Yosef, O., Young, E. D., Tishby, N., & Nelken, I. (2006). Reduction of Information Redundancy in the Ascending Auditory Pathway. *Neuron*, *51*(3), 359–368. https://doi.org/10.1016/j.neuron.2006.06.030

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am*, *118*(2), 887–906.

Choi, I., Lee, J.-Y., & Lee, S.-H. (2018). Bottom-up and top-down modulation of multisensory integration. *Current Opinion in Neurobiology*, *52*, 115–122. https://doi.org/10.1016/j.conb.2018.05.002

Chomsky, N., & Halle, M. (1968). *THE SOUND PATTERN OF ENGLISH*.

Christianson, G. B., Sahani, M., & Linden, J. F. (2008). The Consequences of Response Nonlinearities for Interpretation of Spectrotemporal Receptive Fields. *Journal of Neuroscience*, *28*(2), 446–455. https://doi.org/10.1523/JNEUROSCI.1775-07.2007

David, S. V. (2018). Incorporating behavioral and sensory context into spectro-temporal models of auditory encoding. *Hearing Research*, *360*, 107–123. https://doi.org/10.1016/j.heares.2017.12.021

David, S. V., Fritz, J. B., & Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences*, *109*(6), 2144–2149. https://doi.org/10.1073/pnas.1117717109

David, S. V, Mesgarani, N., Fritz, J. B., & Shamma, S. A. (2009). Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. *Journal of Neuroscience*, *29*(11), 3374–3386.

David, S. V, & Shamma, S. A. (2013). Integration over multiple timescales in primary auditory cortex. *Journal of Neuroscience*, *33*(49), 19154–19166. https://doi.org/10.1523/JNEUROSCI.2270-13.2013

Dean, I., Harper, N. S., & McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience*, *8*(12), 1684–1689. https://doi.org/10.1038/nn1541

Dean, I., Robinson, B. L., Harper, N. S., & McAlpine, D. (2008). Rapid neural adaptation to sound level statistics. *Journal of Neuroscience*, *28*(25), 6430–6438. https://doi.org/10.1523/JNEUROSCI.0470-08.2008

deCharms, R. C., Blake, D. T., & Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science (New York, N.Y.)*, *280*(5368), 1439–1443. https://doi.org/10.1126/science.280.5368.1439

del Castillo, J., & Katz, B. (1954). Statistical factors involved in neuromuscular facilitation and depression. *The Journal of Physiology*, *124*(3), 574–585. https://doi.org/10.1113/jphysiol.1954.sp005130

Depireux, D. A., Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J Neurophysiol*, *85*(3), 1220–1234.

DeWeese, M. R., Hromádka, T., & Zador, A. M. (2005). Reliability and Representational Bandwidth in the Auditory Cortex. *Neuron*, *48*(3), 479–488. https://doi.org/10.1016/j.neuron.2005.10.016

Dimidschstein, J., Chen, Q., Tremblay, R., Rogers, S. L., Saldi, G.-A., Guo, L., Xu, Q., Liu, R., Lu, C., Chu, J., Grimley, J. S., Krostag, A.-R., Kaykas, A., Avery, M. C., Rashid, M. S., Baek,

M., Jacob, A. L., Smith, G. B., Wilson, D. E., … Fishell, G. (2016). A viral strategy for targeting and manipulating interneurons across vertebrate species. *Nature Neuroscience*, *19*(12), 1743–1749. https://doi.org/10.1038/nn.4430

Ding, N., & Simon, J. Z. (2012). 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*, *2012*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Du, J., Blanche, T. J., Harrison, R. R., Lester, H. A., & Masmanidis, S. C. (2011). Multiplexed, High Density Electrophysiology with Nanofabricated Neural Probes. *PLOS ONE*, *6*(10), e26204. https://doi.org/10.1371/journal.pone.0026204

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*, 54–77.

Eggermont, J. J. (2011). Context dependence of spectro-temporal receptive fields with implications for neural coding. *Hearing Research*, *271*(1–2), 123–132. https://doi.org/10.1016/j.heares.2010.01.014

Elgueda, D., Duque, D., Radtke-Schuller, S., Yin, P., David, S. V., Shamma, S. A., & Fritz, J. B. (2019). State-dependent encoding of sound and behavioral meaning in a tertiary region of the ferret auditory cortex. *Nature Neuroscience*, *22*(3), 447–459. https://doi.org/10.1038/s41593-018-0317-8

Englitz, B., David, S., Sorenson, M., & Shamma, S. (2013). MANTA—an open-source, high density electrophysiology recording suite for MATLAB. *Frontiers in Neural Circuits*, *7*. https://www.frontiersin.org/articles/10.3389/fncir.2013.00069

Escabí, M. A., & Read, H. L. (2003). Representation of spectrotemporal sound information in the ascending auditory pathway. *Biological Cybernetics*, *89*(5), 350–362. https://doi.org/10.1007/s00422-003-0440-8

Fettiplace, R. (2020). Diverse Mechanisms of Sound Frequency Discrimination in the Vertebrate Cochlea. *Trends in Neurosciences*, *43*(2), 88–102. https://doi.org/10.1016/j.tins.2019.12.003

Fitzgerald, J. D., Sincich, L. C., & Sharpee, T. O. (2011). Minimal Models of Multidimensional Computations. *PLoS Computational Biology*, *7*(3), e1001111. https://doi.org/10.1371/journal.pcbi.1001111

Fortune, E. S., & Rose, G. J. (2001). Short-term synaptic plasticity as a temporal filter. *Trends in Neurosciences*, *24*(7), 381–385. https://doi.org/10.1016/S0166-2236(00)01835-X

Francis, N. A., Elgueda, D., Englitz, B., Fritz, J. B., & Shamma, S. A. (2018). Laminar profile of task-related plasticity in ferret primary auditory cortex. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-34739-3

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), Article 2. https://doi.org/10.1038/nrn2787

Fritz, J. B., Elhilali, M., David, S. V, & Shamma, S. A. (2007). Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? *Hear Res*, *229*(1–2), 186–203.

Fritz, J. B., Shamma, S. A., Elhilali, M., & Klein, D. J. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci*, *6*(11), 1216–1223.

Froemke, R. C., Merzenich, M. M., & Schreiner, C. E. (2007). A synaptic memory trace for cortical receptive field plasticity. *Nature*, *450*(7168), 425–429. https://doi.org/10.1038/nature06289

Gao, L., Kostlan, K., Wang, Y., & Wang, X. (2016). Distinct Subthreshold Mechanisms Underlying Rate-Coding Principles in Primate Auditory Cortex. *Neuron*, *91*(4), 905–919. https://doi.org/10.1016/j.neuron.2016.07.004

Gao, X., & Wehr, M. (2015). A Coding Transformation for Temporally Structured Sounds within Auditory Cortical Neurons. *Neuron*. https://doi.org/10.1016/j.neuron.2015.03.004

Garofolo, J. S. (1988). *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database.* National Institute of Standards and Technology.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Gruters, K., & Groh, J. (2012). Sounds and beyond: Multisensory and other non-auditory signals in the inferior colliculus. *Frontiers in Neural Circuits*, *6*. https://www.frontiersin.org/articles/10.3389/fncir.2012.00096

Harper, N. S., Schoppe, O., Willmore, B. D. B., Cui, Z., Schnupp, J. W. H., & King, A. J. (2016). Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons. *PLoS Computational Biology*, *12*(11), e1005113. https://doi.org/10.1371/journal.pcbi.1005113

Heffner, H. E., & Heffner, R. S. (1995). Conditioned Avoidance. In G. M. Klump, R. J. Dooling, R. R. Fay, & W. C. Stebbins (Eds.), *Methods in Comparative Psychoacoustics* (pp. 79–93). Birkhauser Verlag.

Heilbron, M., & Chait, M. (2018). Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? *Neuroscience*, *389*, 54–73. https://doi.org/10.1016/j.neuroscience.2017.07.061

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558.

Hsu, A., Borst, A., & Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation and Neural Systems*, *15*, 91–109.

Huang, C. L., & Winer, J. A. (2000). Auditory thalamocortical projections in the cat: Laminar and areal patterns of input. *Journal of Comparative Neurology*, *427*(2), 302–331. https://doi.org/10.1002/1096-9861(20001113)427:2<302::AID-CNE10>3.0.CO;2-J

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106-154.2.

Huetz, C., Gourévitch, B., & Edeline, J. M. (2011). Neural codes in the thalamocortical auditory system: From artificial stimuli to communication sounds. *Hearing Research*, *271*(1–2), 147–158. https://doi.org/10.1016/j.heares.2010.01.010

Hutton, N. (n.d.). *In Silico – a Documentary Film*. In Silico. Retrieved November 18, 2022, from https://insilicofilm.com

Isaacson, J. S., & Scanziani, M. (2011). How Inhibition Shapes Cortical Activity. *Neuron*, *72*(2), 231–243. https://doi.org/10.1016/j.neuron.2011.09.027

Jackman, S. L., & Regehr, W. G. (2017). The Mechanisms and Functions of Synaptic Facilitation. *Neuron*, *94*(3), 447–464. https://doi.org/10.1016/j.neuron.2017.02.047

Joris, P. X., Schreiner, C. E., & Rees, A. (2004). Neural Processing of Amplitude-Modulated Sounds. *Physiological Reviews*, *84*(2), 541–577. https://doi.org/10.1152/physrev.00029.2003

Kato, H. K., Asinof, S. K., & Isaacson, J. S. (2017). Network-Level Control of Frequency Tuning in Auditory Cortex. *Neuron*, *95*(2), 412-423.e4. https://doi.org/10.1016/j.neuron.2017.06.019

Katsiamis, A. G., Drakakis, E. M., & Lyon, R. F. (2007). Practical Gammatone-Like Filters for Auditory Processing. *EURASIP Journal on Audio, Speech, and Music Processing*, *2007*(4), 1–15. https://doi.org/10.1155/2007/63685

Keine, C., Rübsamen, R., & Englitz, B. (2016). Inhibition in the auditory brainstem enhances signal representation and regulates gain in complex acoustic environments. *ELife*, *5*(November 2016). https://doi.org/10.7554/eLife.19295

Keshishian, M., Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2020). Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *ELife*, *9*, e53445. https://doi.org/10.7554/eLife.53445

Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization* (arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980

Klampfl, S., David, S. V, Yin, P., Shamma, S. A., & Maass, W. (2012). A quantitative analysis of information about past and present stimuli encoded by spikes of A1 neurons. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.00935.2011

Klein, D. J., Depireux, D. A., Simon, J. Z., & Shamma, S. A. (2000). Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. *Journal of Computational Neuroscience*, *9*(1), 85–111. https://doi.org/10.1023/a:1008990412183

Knuth, D. E. (2000). *Dancing links* (arXiv:cs/0011047). arXiv. https://doi.org/10.48550/arXiv.cs/0011047

Ko, H., Hofer, S. B., Pichler, B., Buchanan, K. A., Sjöström, P. J., & Mrsic-Flogel, T. D. (2011). Functional specificity of local synaptic connections in neocortical networks. *Nature*, *473*(7345), 87–91. https://doi.org/10.1038/nature09880

Kowalski, N., Depireux, D. A., & Shamma, S. A. (1996). Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *Journal of Neurophysiology*, *76*(5), 3503–3523. https://doi.org/10.1152/jn.1996.76.5.3503

Kozlov, A. S., & Gentner, T. (2016). Central auditory neurons have composite receptive fields. *Proceedings of the National Academy of Sciences*, *113*(5), 1441–1446. https://doi.org/10.1073/pnas.1506903113

Kuchibhotla, K. V, Gill, J. V, Lindsay, G. W., Papadoyannis, E. S., Field, R. E., Sten, T. A. H., Miller, K. D., & Froemke, R. C. (2016). Parallel processing by cortical inhibition enables

context-dependent behavior. *Nature Neuroscience*, *20*(1), 62–71. https://doi.org/10.1038/nn.4436

Kudela, P., Boatman-Reich, D., Beeman, D., Stanley Anderson, W., Carr, C., & Lee Bartlett, E. (2018). Modeling Neural Adaptation in Auditory Cortex. *Frontiers in Neural Circuits*, *12*, 72. https://doi.org/10.3389/fncir.2018.00072

Lakunina, A. A., Menashe, N., & Jaramillo, S. (2022). Contributions of Distinct Auditory Cortical Inhibitory Neuron Types to the Detection of Sounds in Background Noise. *ENeuro*, *9*(2). https://doi.org/10.1523/ENEURO.0264-21.2021

Lakunina, A. A., Nardoci, M. B., Ahmadian, Y., & Jaramillo, S. (2020). Somatostatin-Expressing Interneurons in the Auditory Cortex Mediate Sustained Suppression by Spectral Surround. *Journal of Neuroscience*, *40*(18), 3564–3575. https://doi.org/10.1523/JNEUROSCI.1735-19.2020

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems*, *2*. https://proceedings.neurips.cc/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html

Lesica, N. A., & Grothe, B. (2008a). Efficient temporal processing of naturalistic sounds. *PLoS ONE*, *3*(2), e1655. https://doi.org/10.1371/journal.pone.0001655

Lesica, N. A., & Grothe, B. (2008b). Dynamic Spectrotemporal Feature Selectivity in the Auditory Midbrain. *Journal of Neuroscience*, *28*(21), 5412–5421. https://doi.org/10.1523/JNEUROSCI.0073-08.2008

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), Article 4. https://doi.org/10.1038/nn831

Linden, J. F., & Schreiner, C. E. (2003). Columnar Transformations in Auditory Cortex? A Comparison to Visual and Somatosensory Cortices. *Cerebral Cortex*, *13*(1), 83–89. https://doi.org/10.1093/cercor/13.1.83

Lopez Espejo, M., Schwartz, Z. P., & David, S. V. (2019). Spectral tuning of adaptation supports coding of sensory context in auditory cortex. *PLOS Computational Biology*, *15*(10), e1007430. https://doi.org/10.1371/journal.pcbi.1007430

Lu, K., Liu, W., Zan, P., David, S. V., Fritz, J. B., & Shamma, S. A. (2018a). Implicit Memory for Complex Sounds in Higher Auditory Cortex of the Ferret. *Journal of Neuroscience*, *38*(46), 9955–9966. https://doi.org/10.1523/JNEUROSCI.2118-18.2018

Lu, K., Liu, W., Zan, P., David, S. V., Fritz, J. B., & Shamma, S. A. (2018b). Implicit memory for complex sounds in higher auditory cortex of the ferret. *The Journal of Neuroscience*, 2118–18. https://doi.org/10.1523/JNEUROSCI.2118-18.2018

Lu, T., Liang, L., & Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nature Neuroscience*, *4*(11), Article 11. https://doi.org/10.1038/nn737

Lyall, E. H., Mossing, D. P., Pluta, S. R., Chu, Y. W., Dudai, A., & Adesnik, H. (2021). Synthesis of a comprehensive population code for contextual features in the awake sensory cortex. *ELife*, *10*, e62687. https://doi.org/10.7554/eLife.62687

Ma, Y., Tsao, D., & Shum, H.-Y. (2022). *On the Principles of Parsimony and Self-Consistency for the Emergence of Intelligence* (arXiv:2207.04630). arXiv. https://doi.org/10.48550/arXiv.2207.04630

Maass, W., & Markram, H. (2004). On the computational power of circuits of spiking neurons. *Journal of Computer and System Sciences*, *69*(4), 593–616. https://doi.org/10.1016/j.jcss.2004.04.001

Machens, C. K., Wehr, M. S., & Zador, A. M. (2004). Linearity of cortical receptive fields measured with natural sounds. *Journal of Neuroscience*, *24*(5), 1089–1100.

Malmierca, M. S., Anderson, L. A., & Antunes, F. M. (2015). The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: A potential neuronal correlate for predictive coding. *Frontiers in Systems Neuroscience*, *9*(MAR). Scopus. https://doi.org/10.3389/fnsys.2015.00019

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martinez, L. M., & Alonso, J.-M. (2003). Complex Receptive Fields in Primary Visual Cortex. *The Neuroscientist*, *9*(5), 317–331. https://doi.org/10.1177/1073858403252732

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), Article 9. https://doi.org/10.1038/s41593-018-0209-y

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., Tolias, A. S., Cardin, J. A., & McCormick, D. A. (2015). Waking State: Rapid Variations Modulate Neural and Behavioral Responses. *Neuron*, *87*(6), 1143–1161. https://doi.org/10.1016/j.neuron.2015.09.012

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. https://doi.org/10.1038/nature11020

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science (New York, N.Y.)*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Mesgarani, N., David, S. V, Fritz, J. B., & Shamma, S. A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy*

of Sciences of the United States of America, *111*(18), 6792–6797. https://doi.org/10.1073/pnas.1318017111

Mesgarani, N., Fritz, J. B., & Shamma, S. A. (2010). A computational model of rapid task-related plasticity of auditory cortical receptive fields. *Journal of Computational Neuroscience*, *28*(1), 19–27. https://doi.org/10.1007/s10827-009-0181-3

Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., Oxenham, A. J., Rauschecker, J. P., Tian, B., & Courtenay Wilson, E. (2007). The role of auditory cortex in the formation of auditory streams. *Hearing Research*, *229*(1), 116–131. https://doi.org/10.1016/j.heares.2007.01.007

Miller, L. M., Escabí, M. A., Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal Receptive Fields in the Lemniscal Auditory Thalamus and Cortex. *Journal of Neurophysiology*, *87*(1), 516–527. https://doi.org/10.1152/jn.00395.2001

Miller, L. M., Escabí, M. A., Read, H. L., & Schreiner, C. E. (2001). Functional Convergence of Response Properties in the Auditory Thalamocortical System. *Neuron*, *32*(1), 151–160. https://doi.org/10.1016/S0896-6273(01)00445-7

Montes-Lourido, P., Kar, M., David, S. V., & Sadagopan, S. (2021). Neuronal selectivity to complex vocalization features emerges in the superficial layers of primary auditory cortex. *PLOS Biology*, *19*(6), e3001299. https://doi.org/10.1371/journal.pbio.3001299

Moore, A. K., & Wehr, M. (2013). Parvalbumin-Expressing Inhibitory Interneurons in Auditory Cortex Are Well-Tuned for Frequency. *Journal of Neuroscience*, *33*(34), 13713–13723. https://doi.org/10.1523/JNEUROSCI.0663-13.2013

Moore, R. C., Lee, T., & Theunissen, F. E. (2013). Noise-invariant neurons in the avian auditory cortex: Hearing the song in noise. *PLoS Computational Biology*, *9*(3), e1002942. https://doi.org/10.1371/journal.pcbi.1002942

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain*, *120*(4), 701–722. https://doi.org/10.1093/brain/120.4.701

Murayama, M., Pérez-Garci, E., Nevian, T., Bock, T., Senn, W., & Larkum, M. E. (2009). Dendritic encoding of sensory stimuli controlled by deep cortical interneurons. *Nature*, *457*(7233), Article 7233. https://doi.org/10.1038/nature07663

Nagel, K. I., & Doupe, A. J. (2008). Organizing principles of spectro-temporal encoding in the avian primary auditory area field L. *Neuron*, *58*(6), 938–955.

Natan, R. G., Briguglio, J. J., Mwilambwe-Tshilobo, L., Jones, S. I., Aizenberg, M., Goldberg, E. M., & Geffen, M. N. (2015). Complementary control of sensory adaptation by two types of cortical interneurons. *ELife*, *4*, e09868. https://doi.org/10.7554/eLife.09868

Nelken, I. (2014). Stimulus-specific adaptation and deviance detection in the auditory system: Experiments and models. *Biological Cybernetics*, *108*(5), 655–663. https://doi.org/10.1007/s00422-014-0585-7

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models* (arXiv:2112.10741). arXiv. https://doi.org/10.48550/arXiv.2112.10741

Niwa, M., Johnson, J. S., O'Connor, K. N., & Sutter, M. L. (2012). Active engagement improves primary auditory cortical neurons' ability to discriminate temporal modulation. *Journal of Neuroscience*, *32*(27), 9323–9334. https://doi.org/10.1523/JNEUROSCI.5832-11.2012

Nocon, J. C., Gritton, H. J., James, N. M., Han, X., & Sen, K. (2022). *Parvalbumin neurons, temporal coding, and cortical noise in complex scene analysis* (p. 2021.09.11.459906). bioRxiv. https://doi.org/10.1101/2021.09.11.459906

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, *6*(3), Article 3. https://doi.org/10.1038/s41562-021-01261-y

Oldenburg, I. A., Hendricks, W. D., Handy, G., Shamardani, K., Bounds, H. A., Doiron, B., & Adesnik, H. (2022). *The logic of recurrent circuits in the primary visual cortex* (p. 2022.09.20.508739). bioRxiv. https://doi.org/10.1101/2022.09.20.508739

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), Article 6583. https://doi.org/10.1038/381607a0

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487. https://doi.org/10.1016/j.conb.2004.07.007

Otazu, G. H., Tai, L.-H. H., Yang, Y., & Zador, A. M. (2009). Engaging in an auditory task suppresses responses in auditory cortex. *Nature Neuroscience*, *12*(5), 646–654. https://doi.org/10.1038/nn.2306

Paninski, L., Pillow, J. W., & Simoncelli, E. P. (2004). Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. *Neural Computation*, *16*(12), 2533–2561. https://doi.org/10.1162/0899766042321797

Pennington, J. R., & David, S. V. (2022). *Can Deep Learning Provide a Generalizable Model for Dynamic Sound Encoding in Auditory Cortex?* (p. 2022.06.10.495698). bioRxiv. https://doi.org/10.1101/2022.06.10.495698

Pérez-González, D., & Malmierca, M. S. (2014). Adaptation in the auditory system: An overview. *Frontiers in Integrative Neuroscience*, *8*, 19. https://doi.org/10.3389/fnint.2014.00019

Pi, H.-J., Hangya, B., Kvitsiani, D., Sanders, J. I., Huang, Z. J., & Kepecs, A. (2013). Cortical interneurons that specialize in disinhibitory control. *Nature*, *503*(7477), 521–524. https://doi.org/10.1038/nature12676

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999. https://doi.org/10.1038/nature07140

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, *177*(4), 999-1009.e10. https://doi.org/10.1016/j.cell.2019.04.005

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), Article 7045. https://doi.org/10.1038/nature03687

Rabinowitz, N. C., Willmore, B. D. B., King, A. J., & Schnupp, J. W. H. (2013). Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. *PLOS Biology*, *11*(11), e1001710. https://doi.org/10.1371/journal.pbio.1001710

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

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., & King, A. J. (2012). Spectrotemporal contrast kernels for neurons in primary auditory cortex. *Journal of Neuroscience*, *32*(33), 11271–11284. https://doi.org/10.1523/JNEUROSCI.1715-12.2012

Radtke-Schuller, S., Fritz, J. B., Yin, P., David, S. V, & Shamma, S. A. (2009). A neuroanatomical study of frontal cortical areas in the ferret (Mustela putorius) and their role in top-down control of auditory processing. *3rd International Meeting on Auditory Cortex, Magdeburg, Germany*.

Rahman, M., Willmore, B. D. B., King, A. J., & Harper, N. S. (2019). A dynamic network model of temporal receptive fields in primary auditory cortex. *PLOS Computational Biology*, *15*(5), e1006618. https://doi.org/10.1371/journal.pcbi.1006618

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), Article 1. https://doi.org/10.1038/4580

Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of Complex Sounds in the Macaque Nonprimary Auditory Cortex. *Science*, *268*(5207), 111–114. https://doi.org/10.1126/science.7701330

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms13289

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408. https://doi.org/10.1037/h0042519

Rossant, C., Kadir, S. N., Goodman, D. F. M., Schulman, J., Hunter, M. L. D., Saleem, A. B., Grosmark, A., Belluscio, M., Denfield, G. H., Ecker, A. S., Tolias, A. S., Solomon, S., Buzsáki, G., Carandini, M., & Harris, K. D. (2016). Spike sorting for large, dense electrode arrays. *Nature Neuroscience*, *19*(4), Article 4. https://doi.org/10.1038/nn.4268

Rothman, J. S., Cathala, L., Steuber, V., & Silver, R. A. (2009). Synaptic depression enables neuronal gain control. *Nature*, *457*(7232), 1015–1018. https://doi.org/10.1038/nature07604

Rui, Y. Y., He, J., Zhai, Y. Y., Sun, Z.-H., & Yu, X. J. (2018). Frequency-Dependent Stimulus-Specific Adaptation and Regularity Sensitivity in the Rat Auditory Thalamus. *Neuroscience*, *392*, 13–24. https://doi.org/10.1016/j.neuroscience.2018.09.015

Runyan, C. A., Piasini, E., Panzeri, S., & Harvey, C. D. (2017). Distinct timescales of population coding across cortex. *Nature*, *548*(7665), 92–96. https://doi.org/10.1038/nature23020

Saderi, D., Schwartz, Z. P., Heller, C. R., Pennington, J. R., & David, S. V. (2021). Dissociation of task engagement and arousal effects in auditory cortex and midbrain. *ELife*, *10*, e60153. https://doi.org/10.7554/eLife.60153

Sakata, S., & Harris, K. D. (2009). Laminar Structure of Spontaneous and Sensory-Evoked Population Activity in Auditory Cortex. *Neuron*, *64*(3), 404–418. https://doi.org/10.1016/j.neuron.2009.09.020

Schinkel-Bielefeld, N., David, S. V, Shamma, S. A., & Butts, D. A. (2012). Inferring the role of inhibition in auditory processing of complex natural stimuli. *Journal of Neurophysiology*, *107*(March), 3296–3307. https://doi.org/10.1152/jn.01173.2011

Schneider, D., Audette, N., & Zhou, W. (2021). Expectation of self-generated sounds drives predictive processing in mouse auditory cortex. *The Journal of the Acoustical Society of America*, *150*(4), A106–A106. https://doi.org/10.1121/10.0007779

Schwartz, Z. P., Buran, B. N., & David, S. V. (2020). Pupil-associated states modulate excitability but not stimulus selectivity in primary auditory cortex. *Journal of Neurophysiology*, *123*(1), 191–208. https://doi.org/10.1152/jn.00595.2019

Schwartz, Z. P., & David, S. V. (2018). Focal suppression of distractor sounds by selective attention in auditory cortex. *Cerebral Cortex (New York, N.Y. : 1991)*, *28*(1), 323–339. https://doi.org/10.1093/cercor/bhx288

Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and Statistical Modeling with Python*. 92–96. https://doi.org/10.25080/Majora-92bf1922-011

Seay, M. J., Natan, R. G., Geffen, M. N., & Buonomano, D. V. (2020). Differential Short-Term Plasticity of PV and SST Neurons Accounts for Adaptation and Facilitation of Cortical Neurons to Auditory Tones. *Journal of Neuroscience*, *40*(48), 9224–9235. https://doi.org/10.1523/JNEUROSCI.0686-20.2020

See, J. Z., Atencio, C. A., Sohal, V. S., & Schreiner, C. E. (2018). Coordinated neuronal ensembles in primary auditory cortical columns. *ELife*, *7*. https://doi.org/10.7554/eLife.35587

Sen, K., Theunissen, F. E., & Doupe, A. J. (2001). Feature Analysis of Natural Sounds in the Songbird Auditory Forebrain. *Journal of Neurophysiology*, *86*(3), 1445–1458. https://doi.org/10.1152/jn.2001.86.3.1445

Shamma, S. A., Fleshman, J. W., Wiser, P. R., & Versnel, H. (1993). Organization of response areas in ferret primary auditory cortex. *Journal of Neurophysiology*, *69*(2), 367–383.

Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current Opinion in Neurobiology*, *21*(5), 761–767. https://doi.org/10.1016/j.conb.2011.05.027

Sibille, J., Gehr, C., Benichov, J. I., Balasubramanian, H., Teh, K. L., Lupashina, T., Vallentin, D., & Kremkow, J. (2022). High-density electrode recordings reveal strong and specific connections between retinal ganglion cells and midbrain neurons. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-32775-2

Siegle, J. H., López, A. C., Patel, Y. A., Abramov, K., Ohayon, S., & Voigts, J. (2017). Open Ephys: An open-source, plugin-based platform for multichannel electrophysiology. *Journal of Neural Engineering*, *14*(4), 045003. https://doi.org/10.1088/1741-2552/aa5eea

Silver, R. A. (2010). Neuronal arithmetic. *Nature Reviews Neuroscience*, *11*(7), Article 7. https://doi.org/10.1038/nrn2864

Simon, J. Z., Depireux, D. A., Klein, D. J., Fritz, J. B., & Shamma, S. A. (2007). Temporal symmetry in primary auditory cortex: Implications for cortical connectivity. *Neural Computation*, *19*(3), 583–638. https://doi.org/10.1162/neco.2007.19.3.583

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *Journal of the Acoustical Society of America*, *114*(6 Pt 1), 3394–3411.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., & Morcos, A. S. (2022). *Beyond neural scaling laws: Beating power law scaling via data pruning* (arXiv:2206.14486). arXiv. https://doi.org/10.48550/arXiv.2206.14486

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Chen, S., Colonell, J., Gardner, R. J., Karsh, B., Kloosterman, F., Kostadinov, D., Mora-Lopez, C., O'Callaghan, J., Park, J., … Harris, T. D. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, *372*(6539), eabf4588. https://doi.org/10.1126/science.abf4588

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, *571*(7765), Article 7765. https://doi.org/10.1038/s41586-019-1346-5

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., & Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, *364*(6437), eaav7893. https://doi.org/10.1126/science.aav7893

Studer, F., & Barkat, T. R. (2022). Inhibition in the auditory cortex. *Neuroscience & Biobehavioral Reviews*, *132*, 61–75. https://doi.org/10.1016/j.neubiorev.2021.11.021

Syeda, A., Zhong, L., Tung, R., Long, W., Pachitariu, M., & Stringer, C. (2022). *Facemap: A framework for modeling neural activity based on orofacial tracking* (p. 2022.11.03.515121). bioRxiv. https://doi.org/10.1101/2022.11.03.515121

Tan, Z., Hu, H., Huang, Z. J., & Agmon, A. (2008). Robust but delayed thalamocortical activation of dendritic-targeting inhibitory interneurons. *Proceedings of the National Academy of Sciences*, *105*(6), 2187–2192. https://doi.org/10.1073/pnas.0710628105

Theunissen, F. E., David, S. V, Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatial temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, *12*, 289–316.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds. *Journal of Neuroscience*, *20*(6), 2315–2331. https://doi.org/10.1523/JNEUROSCI.20-06-02315.2000

Thorson, I. L., Liénard, J., & David, S. V. (2015). The Essential Complexity of Auditory Receptive Fields. *PLOS Computational Biology*, *11*(12), e1004628. https://doi.org/10.1371/journal.pcbi.1004628

Tischbirek, C. H., Noda, T., Tohmi, M., Birkner, A., Nelken, I., & Konnerth, A. (2019). In Vivo Functional Mapping of a Cortical Column at Single-Neuron Resolution. *Cell Reports*, *27*(5), 1319-1326.e5. https://doi.org/10.1016/j.celrep.2019.04.007

Trainito, C., von Nicolai, C., Miller, E. K., & Siegel, M. (2019). Extracellular Spike Waveform Dissociates Four Functionally Distinct Cell Classes in Primate Cortex. *Current Biology*, *29*(18), 2973-2982.e5. https://doi.org/10.1016/j.cub.2019.07.051

Trussell, L. O. (1999). Synaptic Mechanisms for Coding Timing in Auditory Neurons. *Annual Review of Physiology*, *61*(1), 477–496. https://doi.org/10.1146/annurev.physiol.61.1.477

Tsodyks, M., Pawelzik, K., & Markram, H. (1998). Neural networks with dynamic synapses. *Neural Computation*, *10*(4), 821–835.

Ulanovsky, N., Las, L., Farkas, D., & Nelken, I. (2004). Multiple time scales of adaptation in auditory cortex neurons. *Journal of Neuroscience*, *24*(46), 10440–10453. https://doi.org/10.1523/JNEUROSCI.1905-04.2004

Ulanovsky, N., Las, L., & Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, *6*(4), Article 4. https://doi.org/10.1038/nn1032

van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., & Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences*, *111*(40), 14332–14341. https://doi.org/10.1073/pnas.1402773111

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, *22*(12), Article 12. https://doi.org/10.1038/s41593-019-0517-x

Wallace, M. N., Roeda, D., & Harper, M. S. (1997). Deoxyglucose uptake in the ferret auditory cortex. *Experimental Brain Research*, *117*(3), 488–500. https://doi.org/10.1007/s002210050245

Watkins, P. V, & Barbour, D. L. (2011). Rate-level responses in awake marmoset auditory cortex. *Hearing Research*, *275*(1–2), 30–42. https://doi.org/10.1016/j.heares.2010.11.011

Wehr, M., & Zador, A. M. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature*, *426*(6965), Article 6965. https://doi.org/10.1038/nature02116

Weiss, A. R., Liguore, W. A., Domire, J. S., Button, D., & McBride, J. L. (2020). Intra-striatal AAV2.retro administration leads to extensive retrograde transport in the rhesus macaque brain: Implications for disease modeling and therapeutic development. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-63559-7

Whitmire, C. J., & Stanley, G. B. (2016). Rapid Sensory Adaptation Redux: A Circuit Perspective. *Neuron*, *92*(2), 298–315. https://doi.org/10.1016/j.neuron.2016.09.046

Williams, A. (2022). *The Exploited Labor Behind Artificial Intelligence*. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence

Williamson, R. S., Ahrens, M. B., Linden, J. F., & Sahani, M. (2016). Input-Specific Gain Modulation by Local Sensory Context Shapes Cortical and Thalamic Responses to Complex Sounds. *Neuron*, *91*(2), 467–481. https://doi.org/10.1016/j.neuron.2016.05.041

Willmore, B. D. B., Schoppe, O., King, A. J., Schnupp, J. W. H., & Harper, N. S. (2016). Incorporating Midbrain Adaptation to Mean Sound Level Improves Models of Auditory Cortical Processing. *Journal of Neuroscience*, *36*(2), 280–289. https://doi.org/10.1523/JNEUROSCI.2441-15.2016

Winer, J. A., Miller, L. M., Lee, C. C., & Schreiner, C. E. (2005). Auditory thalamocortical transformation: Structure and function. *Trends in Neurosciences*, *28*(5), 255–263. https://doi.org/10.1016/j.tins.2005.03.009

Wu, M. C.-K., David, S. V, & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*, 477–505.

Yarden, T. S., Mizrahi, A., & Nelken, I. (2022). Context-Dependent Inhibitory Control of Stimulus-Specific Adaptation. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0988-21.2022

Yarden, T. S., & Nelken, I. (2017). Stimulus-specific adaptation in a recurrent network model of primary auditory cortex. *PLOS Computational Biology*, *13*(3), e1005437. https://doi.org/10.1371/journal.pcbi.1005437

Yaron, A., Hershenhoren, I., & Nelken, I. (2012). Sensitivity to complex statistical regularities in rat auditory cortex. *Neuron*, *76*(3), 603–615. https://doi.org/10.1016/j.neuron.2012.08.025

Yin, P., Fritz, J. B., & Shamma, S. A. (2010). Do ferrets perceive relative pitch? *J Acoust Soc Am*, *127*(3), 1673–1680.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, *31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Zhang, Q., Hu, X., Hong, B., & Zhang, B. (2019). A hierarchical sparse coding model predicts acoustic feature encoding in both auditory midbrain and cortex. *PLOS Computational Biology*, *15*(2), e1006766. https://doi.org/10.1371/journal.pcbi.1006766