Chapter 3

Bottom-up characterization: single-cell level analysis

How does the brain work? It is in fact more of a philosophy what kind of answer one would consider satisfactory; it could be full descriptions of gene expression profiles and the interactions, it could be entire wiring diagrams of the brain, or it could be complete mathematical formulations that govern spiking patterns of neurons. An answer I pursue in this chapter belongs to the latter class; i.e., the simplest model that can account for neural responses to sensory stimuli—specifically, the one describing how neurons in the auditory cortex behave in the presence of acoustic stimuli. Because our knowledge on the response dynamics is limited, however, this chapter explores bottom-up experimental characterizations of neural behaviors at the single-cell level to obtain necessary information for building a plausible model well-tailored to the observed properties. We used in vivo whole-cell patch-clamp recordings and assessed how neural responses in the (anesthetized rat) auditory cortex depend on stimulus history and its context (Asari and Zador, in preparation).

This chapter consists of six sections. Section 3.1 first gives a brief overview on previous approaches for modeling neural responses (Section 3.1.1), and describes motivations as well as background studies on context-dependence in the auditory systems (Section 3.1.2).
A novel experimental approach is then described in Section 3.2 for estimating the time course of context-dependence of auditory neurons. Details of data analysis methods are described in Section 3.3, and some neural encoding models employed here—such as spectro-temporal receptive field (STRF) models—in Section 3.4 (see also Appendix Sections A.2–A.3). Section 3.5 shows our main results; i.e., the context-dependent effects generally decayed over seconds, and could last as long as four seconds or longer in some neurons (Section 3.5.1). Such effects were shown to be more susceptible to the changes in lower-order sound properties such as overall stimulus intensities than to the changes in higher-order properties such as amplitude-modulation rates (Section 3.5.2). We also estimated the upper bound for the performance of an encoding model when only a finite length of stimulus history is available (Section 3.5.3), and identified that a window length of at least several seconds was required to capture the stimulus-related predictable response power fully enough. As shown in Section 3.5.4, however, the linear model performance did not improve substantially by just extending the window length and even by incorporating static nonlinearities. Finally, in Section 3.6, we will discuss relations to previous studies (Sections 3.6.1–3.6.2) as well as plausible mechanisms and roles of context-dependence in auditory signal processing (Section 3.6.3)—e.g., in stream segregation and others that require integration over seconds—and speculate possible directions to improve encoding models for the auditory cortical neurons (Section 3.6.4).

3.1 Introduction

Scientists have typically taken two complementary approaches to elucidate the laws of nature; observation and theory. To understand the sensory coding mechanisms in the brain for example, we first look for “neural correlates” of the signal processing (often those of psychophysical counterparts), and then rationalize the observations under some assumptions or hypotheses, often in mathematical terms to build up a theory. Such a theory might in turn be supported or falsified by further experimental observations, and these processes of “trials and errors” are the
key basic steps we could take towards a full understanding of a complex system such as the nervous system.

Any statement that describes a system is a model, be it qualitative or quantitative; we make some assumptions and give a most reasonable description and interpretation for empirical or theoretical understandings. Here I focus on quantitative ones—we cannot do without equations, I believe, to completely clarify a system—and Section 3.1.1 briefly overviews existing sensory coding models on single-neuron dynamics. Such quantitative models in neuroscience could range over various levels—in one extreme we can build models on the basis of detailed biophysical properties of ion channels and cell membranes (Hodgkin and Huxley, 1952; FitzHugh, 1961; Nagumo et al., 1964; Gerstner and Kistler, 2002; Agüera y Arcas and Fairhall, 2003; Agüera y Arcas et al., 2003), and in the other extreme we could treat the system as a “black-box” (Herz et al., 2006)—but those in sensory systems neuroscience can be typically classified into the latter type in a broad sense.

Section 3.1.2 then concisely describes the “observations” reported in previous work, and describes what has been missing for building a plausible encoding model—here we focus on the relevant time-scales of response dynamics in the auditory cortex to identify the minimum necessary window length for a neural encoding model.

### 3.1.1 Sensory Coding Models

One goal of neuroscience is to understand how the brain processes sensory information. As a first step, computational and systems neuroscience often targets to characterize the relationship between inputs (sensory stimuli) and outputs (neural responses), and a plenty of models and methods on this topic have been proposed over the past decades (for review, see e.g., Meister and Berry, 1999; Simoncelli et al., 2004; Herz et al., 2006; Wu et al., 2006). Such sensory coding models typically describe either of two complementary aspects of the sensory signal processing in the brain. The first—and best studied—is the encoding process by which a stimulus is converted by the nervous system into neural activity. Less studied is the decoding process,
where experimenters attempt to use neural activity to reconstruct the stimulus that evoked it. Because decoding models typically require considering the activity of populations of neurons (e.g., linear generative models for sparse or efficient codings; Olshausen and Field, 1996; van Hateren and Ruderman, 1998; Lewicki, 2002; Smith and Lewicki, 2005, 2006; see also Chapter 2), they are less experimentally accessible (but see e.g., Bialek et al., 1991; Warland et al., 1997; Stanley et al., 1999) and thus here we focus on encoding directions on single-neuron dynamics.

The central ideas of both classical and recent—more sophisticated—encoding models are; (1) to extract features as linear combinations of the inputs; and (2) to model the outputs as a nonlinear function of these features. These two steps can be considered as a “unit” process in artificial neural networks, and many encoding models—often consisting of cascades of linear-nonlinear transformations—can then be thought of as multi-layer neural networks with fewer “hidden” units to achieve sufficient dimensionality reduction for interpretation purposes (see also Appendix Section A.3.2). In this view, many learning methods developed separately in different fields including neuroscience can be considered as identical. In fact, a goal is essentially the same between the fields; “pattern” or feature identification in the data of interest (Victor, 2005). What is critical for sensory coding models is then not the (re)discovery of the methods but rather the interpretation from biological viewpoints. I would put forward that, with biologically reasonable interpretations, we could apply state-of-the-art techniques in different fields (e.g., statistics or machine learning; Tarca et al., 2007) for our own purposes, such as to decipher the neural code. Appropriate interpretations would then provide some insights on the fundamental questions of how we can go further from the characterizations to the understandings of the mechanisms underlying neural behaviors, or how we can go further from “neural correlates” to their functions (see also Chapter 4).

A good realistic model should then have both high expressive power—i.e., good fitness to data—and biologically reasonable interpretations. However, there is often a trade-off between them. On the one hand, artificial neural networks have in general high expressive power
(Hornik et al., 1989; Barron, 1993) but they are not so popular in neuroscience because the relation between model parameters and experimentally measured quantities is obscure (but see Lehky et al., 1992; Lau et al., 2002; Prenger et al., 2004). On the other hand, linear models have been widely used in the auditory—and visual—physiology due to their simplicity and interpretability (Eggermont et al., 1983; Klein et al., 2000; Theunissen et al., 2001; Escabí and Schreiner, 2002; Linden et al., 2003; Machens et al., 2004). Linear spectro-temporal receptive field (STRF) models have been fairly successful in describing the input-output function of some stimulus ensembles in auditory cortex (Kowalski et al., 1996; Depireux et al., 2001), but have generally yielded only poor results for other ensembles, including those consisting of natural stimuli (Machens et al., 2004). In this thesis, I began with the simplest model—i.e., the linear STRF model—because the concept of “receptive field” (Barlow, 1953; Kuffler, 1953; Ringach, 2004) is so simple and easy to interpret, and aimed at (1) experimentally collecting relevant information on neural dynamics and (2) searching directions to improve the model performance based on the experimental results.

### 3.1.2 Motivation and Background Studies

Why has the classical STRF-based approach failed to provide a general model? A straightforward answer is that the actual input-output function is nonlinear; e.g., it might include multiplicative interactions between different frequency bands. However, the space of nonlinear functions is huge, and it is not feasible to fit general high-order models (but see Marmarelis and Naka, 1972; Eggermont, 1993; Rotman et al., 2001); e.g., if the input spectrogram is discretized with a (rather coarse) frequency resolution of 1/4 octave over 5 octaves and a (rather coarse) temporal resolution of 10 msec over 200 msec, then the number of parameters for a linear model is \( N = \frac{5 \times 4}{1} \times \frac{200}{10} = 400 \), while \( \mathcal{O}(N^2) \) for a second-order Wiener model, and in general \( \mathcal{O}(N^n) \) for an \( n \)-th order Wiener model (Volterra, 1930; Wiener, 1958). Thus the success of general “black-box” models is quickly limited by the “curse of dimensionality” (Bellman, 1961), i.e., the fact that the amount of data required to fit a general model increases
exponentially with the order of the model; even general black-box models—such as multi-layer artificial neural networks—that are guaranteed to succeed in principle are often data-limited in practice. Although in some cases these difficulties can be circumvented by the judicious choice of nonlinearities (e.g., Chichilnisky, 2001; Schwartz and Simoncelli, 2001; Fishbach et al., 2001, 2003; Nykamp and Ringach, 2002; Paninski, 2003a; Sharpee et al., 2004, 2006; Rust et al., 2005), it is difficult to know \textit{a priori} what form the nonlinearities should take.

One way to reduce the number of model parameters is to tailor the model to the observed properties of auditory cortex neurons. The parameter count above illustrates that the system’s “memory”—i.e., the dependence of the neuron’s input-output behavior on stimulus history or context—is one of the primary determinants of model complexity; doubling the length of the memory (e.g., from 200 to 400 msec in the example above) doubles the number of input variables (from \(N\) to \(2N\) for fixed temporal resolution). Thus it would be useful to characterize the length of the system’s memory.

Sensory signal processing in the brain indeed depends on stimulus history and its context. The visual system is by far the best characterized sensory system, and the effects of stimulus history have been examined over a wide range of time-scales, from tens of milliseconds to minutes (for a review, see Kohn, 2007). For acoustic signal processing, psychophysical studies have demonstrated that stimulus context strongly affects sound perception and auditory scene analysis in humans (Massaro, 1972; Bregman, 1990; Oxenham, 2001), and physiological studies on the sensitivities to stimulus context (e.g., forward masking) suggest that the responses of neurons in the primary auditory cortex (area A1) are highly dependent on stimulus history (e.g., Abeles and Goldstein, 1972; Hocherman and Gilat, 1981; Phillips, 1985; Calford and Semple, 1995; Brosch and Schreiner, 1997, 2000; Wehr and Zador, 2005). Most auditory physiological studies so far, however, have looked only at shorter time-scales (i.e., tens or hundreds of milliseconds) except for some recent works (Condon and Weinberger, 1991; Malone et al., 2002; Ulanovsky et al., 2003, 2004; Bartlett and Wang, 2005), and it remains to be addressed how long and how large past events would influence the activity of neurons in A1.
We have therefore developed a novel experimental approach for estimating the time course of context-dependence of neurons in the primary auditory cortex of anesthetized rats and compared it to several encoding models (Sections 3.2–3.4). Whole-cell patch-clamp recordings in vivo were used to examine subthreshold responses to various sequences of natural and synthetic stimulus ensembles with rich temporal and spectral structures. We found that context-dependence could last for a surprisingly long time—sometimes as long as four seconds or longer (Section 3.5). However, extending the memory of linear models did not improve their performance, indicating that these long-lasting effects of context were nonlinear. The slow stimulus adaptation may play a role in stream segregation and other forms of auditory processing that require integration over seconds (Section 3.6).

3.2 Experimental Methods

3.2.1 Surgery

Long-Evans rats (20–27 days old) were anesthetized (30 mg/kg ketamine and 0.24 mg/kg medetomidine) in strict accordance with the National Institutes of Health guidelines as approved by the Cold Spring Harbor Laboratory Animal Care and Use Committee. After the animal was deeply anesthetized, it was placed in a custom naso-orbital restraint, which left the ears free and clear. A cisternal drain was made, and a small craniotomy and durotomy were performed above the left primary auditory cortex (area A1). The cortex was covered with physiological buffer (in mM: NaCl, 127; Na₂CO₃, 25; NaH₂PO₄, 1.25; KCl, 2.5; MgCl₂, 1; and glucose, 25) mixed with 1.5% agarose. Temperature was monitored rectally and maintained at 37 °C using a feedback controlled blanket. Depth of anesthesia was monitored throughout the experiment, and supplemental anesthesia was provided when required.
3.2.2 Whole-Cell Patch-Clamp Recordings

Whole-cell recordings were obtained \( \textit{in vivo} \) using standard blind patch-clamp methods (Pei et al., 1991; Margrie et al., 2002; Machens et al., 2004). Electrodes were pulled from filamented, thin-walled, borosilicate glass (outer diameter, 1.5 mm; inner diameter, 1.17 mm; World Precision Instruments, Sarasota, FL) on a vertical two-stage puller (Narishige, East Meadow, NY). Internal solution contained (in mM): KCl, 20; K-Gluconate, 100; HEPES, 10; MgCl\(_2\), 2; CaCl\(_2\), 0.05; Mg-ATP, 4; Na\(_2\)-GTP, 0.3; Na\(_2\)-Phosphocreatine, 10; and \( \sim 2.5\% \) micro-emerald or micro-ruby (dextran-conjugated fluorescent dye; Invitrogen, catalogue number: D-7156 and D-7162, respectively); pH 7.3; diluted to 275 mOsm. Resistance to bath was 3.5–5.0 MΩ before seal formation. We used a custom data acquisition system written in MATLAB (Mathworks, Natick, MA), and sampled membrane potential at 10 kHz using an Axopatch 200B (Molecular Devices, Palo Alto, CA) in current-clamp mode with no online series resistance compensation. Mean series resistance was \( 68.8 \pm 16.7 \) MΩ (mean ± standard deviation; 189 cells). Note that whole-cell recordings select for neurons solely on the basis of the experimenter’s ability to form a “GΩ seal,” as opposed to conventional extracellular methods (e.g., using tungsten electrodes; Hubel, 1957) that have a selection bias towards high-firing neurons.

Recordings were made from A1 as determined by the tonotopic gradient and by the frequency-amplitude tuning properties of cells and local field potentials. We recorded from almost all subpial depths (range: 85–847 µm, as determined from micro-manipulator travel; see also Figure 3.6). Thirteen cells were recovered histologically, which were verified to be pyramidal cells (e.g., Figure 3.4C). All together, we recorded from 194 cells in 139 animals, out of which 123 cells met our criterion for the analysis (see below Section 3.3). Of these, 39 cells were examined with natural sounds, 14 cells with temporally-orthogonal ripple combinations, 6 cells with dynamic moving ripples, 39 cells with modulated harmonic tones, and 27 cells with modulated pink noise (see below Section 3.2.4, and also Table 3.1 and Figure 3.2).
Sound properties varied among conditioning stimuli

<table>
<thead>
<tr>
<th>Probe</th>
<th>All</th>
<th>Amplitude</th>
<th>Frequency</th>
<th>AM</th>
<th>FM</th>
<th>Higher-order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Sound</td>
<td>305 (39)</td>
<td>23 (23)</td>
<td>25 (25)</td>
<td>25 (25)</td>
<td>25 (25)</td>
<td>63 (27)</td>
</tr>
<tr>
<td>TORC</td>
<td>39 (9)</td>
<td>40 (9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMR</td>
<td>8 (2)</td>
<td>20 (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHT</td>
<td>25 (20)</td>
<td>25 (21)</td>
<td></td>
<td>71 (27)</td>
<td>57 (25)</td>
<td></td>
</tr>
<tr>
<td>MPN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>74 (27)</td>
</tr>
</tbody>
</table>

Table 3.1: **Summary of recording data.** Shown is the number of probe stimuli tested in at least two different contexts repeated over at least four trials. Each row represents the probe stimulus type whereas each column indicates the stimulus properties varied among the conditioning stimuli. The corresponding number of recorded cells in A1 is shown in parentheses. A given cell could be tested with more than one probe stimulus, each of which could in turn be tested with more than one type of conditioning stimulus ensemble (see also Sections 3.2.3–3.2.4 and Figure 3.1). **Natural sounds** would differ in all possible sound properties, allowing us to examine overall context-dependence effects (Figures 3.4 and 3.5), whereas synthetic sounds (**TORC, DMR, MHT, and MPN**) were used to examine the effects caused by the changes in each of the following sound properties among conditioning stimuli (Figure 3.8); **amplitude** over the maximum range of 60 dB attenuation, **frequency** with the maximum shift of 4 octaves (e.g., Figure 3.7), **AM** and **FM** with the maximum difference of 20 Hz (rate) and 3-fold (depth), and **higher-order** properties by comparing natural sounds and corresponding modulated pink noise. Acronyms: TORC, temporally-orthogonal ripple combination; DMR, dynamic moving ripple; MHT, modulated harmonic tone; MPN, modulated pink noise; AM, amplitude-modulation; FM, frequency-modulation; A1, primary auditory cortex.
3.2.3 Stimulus Design

During the recordings, sequences of various stimulus combinations were presented in a randomly interleaved manner. To maximize the yield in finite recording length (typically \( \sim 20 \) min), \( N \) sequences were generated for \( N \) stimulus fragments (\( S_i \) for \( i = 1, \ldots, N \)) that allow us to examine the responses to all stimulus pairs (\( S_i S_j \); conditioning stimulus \( S_i \), probe stimulus \( S_j \)) and to each stimulus following a “silent” period, i.e., an inter-sequence interval. More formally, we presented given \( N \) stimulus fragments and a “silent” period as in a cyclic code over a finite field \( \mathbb{F}_{N+1} \) of block length 2 (van Lint, 1992). For \( N = 4 \), for example, we would randomly present the following four sequences: \( S_1 S_3 S_2 S_4, S_3 S_4 S_1 S_2, S_4 S_2 S_3 S_5 S_1 \), and \( S_2 S_2 S_1 S_4 S_3 \) (Figure 3.1A). Inter-stimulus intervals and inter-sequence intervals were 0 and \( \sim 6 \) sec, respectively.

Natural sound ensembles were used to examine the overall context-dependent effects (Section 3.5.1), whereas synthetic sound ensembles (with or without one additional natural sound fragment) were used to examine the effects of the changes in each of the following acoustic properties (Section 3.5.2); amplitude, frequency, amplitude-modulation (AM), frequency-modulation (FM), and higher-order spectro-temporal acoustic features (for details of the stimulus ensembles, see Section 3.2.4 and also Figure 3.1B). Although here we used the same stimulus design in these two cases, care should be taken for the analysis. Any stimulus fragment can be a probe stimulus in both cases, and in the former analysis, any stimulus (including “silence”) can be a conditioning stimulus as well (see also Figures 3.1A and 3.4). In the latter case, however, only an appropriate set of stimulus fragments can be used as conditioning stimuli; e.g., to assess the frequency change effects with three synthetic frequency variants (\( S_i \) for \( i = 1, \ldots, 3 \)) and one natural sound fragment (\( S_4 \)), the stimulus design in Figure 3.1A for instance can be used for the recordings, but the analysis should be conducted by using only those frequency variants (\( S_i \) for \( i = 1, \ldots, 3 \)) as conditioning stimuli for each probe fragment (\( S_i \) for \( i = 1, \ldots, 4 \); see also Figure 3.7). For details of the analysis, see Section 3.3.
Figure 3.1: Experimental design and auditory stimuli used in this study. (A) Experimental design for analyzing context-dependence. During the recording, we presented well-designed $N$ sequences of a given set of $N$ sound fragments with no inter-stimulus interval in a randomly interleaved manner (shown is an example for $N = 4$; inter-sequence interval was $\sim$6 sec in this study). For the analysis, we aligned the recording data to examine the response variability to a given sound fragment (probe; $S_1$ in this example) due to the presence of different preceding stimuli (context; “silence,” $S_1$, $S_2$, $S_3$, and $S_4$ in this example). The choice of conditioning stimuli depends on the goal of the analysis (see Sections 3.2.3–3.2.4). Note that the response power (light gray line; $P[\nu_{ij}(t)]$ from Eq.(3.12)) to a probe stimulus at time $t$ from the onset of the probe can be divided into noise power ($P[\varepsilon_{ij}(t)]$) and stimulus-related power (black line; $P[\mu(t) + \nu(t)]$ in Eq.(3.18)) that can be further decomposed into context-independent fraction ($P[\mu(t)]$ in Eq.(3.19)) and context-dependent fraction (dark gray line; $P[\nu(t)]$ in Eq.(3.20)).

(B) Natural sounds and synthetic sounds. Natural sound fragments ($S_{NS1}$ and $S_{NS2}$; 4.11 sec long; sound pressure waveforms, spectrograms, and temporal and spectral marginal distributions are shown.) differ a lot, which causes a large and long context-dependence when they are used as conditioning stimuli (Figures 3.4 and 3.5). On the other hand, the temporal and spectral patterns in the marginal distributions between modulated pink noise $S_{MPN1}$ and the corresponding natural sound $S_{NS1}$ are nearly identical, resulting in a small and short context-dependence (Figures 3.8 and 3.9). Synthetic sounds such as modulated harmonic tones (Eq.(3.4); sound pressure waveform and corresponding spectrograms; 1 sec long) can be used to assess the effects of the changes in sound properties in more detail. Compared to $S_{MHT}$ for example, $S_{\Delta AMP}$ has 30 dB less power, $S_{\Delta FREQ}$ has the frequency components up-shifted by 1.5 octaves, $S_{\Delta AM}$ has slower amplitude-modulation rates by 4 Hz on average, and $S_{\Delta FM}$ has half the standard deviation for the frequency-modulation depth. We also used temporally-orthogonal ripple combinations and dynamic moving ripples (Eqs.(3.1)–(3.3); examples are not shown here, but see elsewhere, e.g., Klein et al., 2000; Escabí and Schreiner, 2002) but only for examining the effects of amplitude and frequency changes in conditioning stimuli. For details, see Sections 3.2–3.3.
3.2.4 Stimuli

All stimuli were delivered at 97.656 kHz using a TDT System 3 with an ED1 electrostatic speaker (Tucker-Davis Technologies, Alachua, FL) in free-field configuration (speaker located ~8 cm lateral to, and facing, the contralateral ear) in a double-walled sound booth (Industrial Acoustics Company, Bronx, NY). The speaker had a maximum intensity (at 10 V command voltage) of 92 dB sound pressure level (SPL), and its frequency response was flat from 1 to 22 kHz to within standard deviation of 3.7 dB. Sound levels were measured with a type 7012 one-half inch ACO Pacific microphone (ACO Pacific, Belmont, CA) positioned where the contralateral ear would be (but with the absence of animal).

Natural Sounds

All natural sound fragments were taken from commercially available audio compact discs (CDs), originally sampled at 44.1 kHz and resampled at 97.656 kHz for stimulus presentation: The Diversity of Animal Sounds and Sounds of Neotropical Rainforest Mammals (Cornell Laboratory of Ornithology, Ithaca, NY). The majority of the sound sections lasted for 3.5–6.5 sec, but some were shorter (2–3 sec) to examine as many stimulus combinations as possible (see above the cyclic stimulus design in Section 3.2.3 and Figure 3.1). The sound segments were chosen from original sound tracks to have minimum “silent” periods (especially at the onset and termination), and a 5 msec cosine-squared ramp was applied at the onset and termination to make sure a smooth connection between the segments even with no inter-stimulus interval. The peak amplitude of each segment was normalized to the ±10 V range of the speaker driver. The natural sound stimuli consisted of 46 different sounds in total, covered almost all frequencies from 0 to 22 kHz, and ranged from narrow- to broad-band stimuli. Note however that only a subset of stimuli (typically N~7 fragments) was tested on any particular cell.

---

1No “click” sound (due to discontinuity) was detected at the connection point at least to my ears.
Synthetic Sounds

All synthetic sounds were sampled at 97.656 kHz and lasted for 4–5.5 sec. We used temporally-orthogonal ripple combinations (TORCs; for details, see e.g., Klein et al., 2000) and dynamic moving ripples (DMRs; for details, see e.g., Escáibá and Schreiner, 2002) to examine how the changes in amplitudes and frequencies in conditioning stimuli would affect the responses to following probe stimuli. Modulated harmonic tones were used to assess the effects of AM and FM changes as well as the changes in amplitudes and frequencies; and modulated pink noise was used for testing the effects of the changes in higher-order sound properties, such as complex interactions between spectro-temporal constituents (Figure 3.1B and Table 3.1). In this study, no cell was tested with all the sound properties due to a limited recording length (typically, \(\sim\) 20 min). Strictly speaking, we cannot directly compare the context-dependence effects caused by the changes in different acoustic properties; however, the comparisons we made (Figure 3.9) would not be unfair because the acoustic properties were varied across almost entire range that A1 neurons can follow faithfully (e.g., sound trains or temporal modulations up to tens of Hz; Creutzfeldt et al., 1980; Joris et al., 2004).

**Temporally-Orthogonal Ripple Combinations** The following equation was used to generate ripples and their combinations:

\[
y(t) = \sum_i y_E(t, x_i) \times y_C(t, f_i),
\]  

(3.1)

where the envelope \(y_E(t, x_i)\) and carriers \(y_C(t, f_i)\) are respectively given by:

\[
20 \log_{10}[y_E(t, x_i)] = a_0 + \sum_{j,k} \frac{a_{jk}}{2} \cos[2\pi(\psi_j(t) + \Omega_k(t) x_i) + \phi_{jk}],
\]  

(3.2)

\[
y_C(t, f_i) = \sin[2\pi f_i t + \varphi_i].
\]  

(3.3)
Note that $a_{k,j} (> 0)$ is a sinusoidal modulation depth around the mean $a_0$ in dB, $\omega_j(t) = \partial \Psi_j / \partial t$ in Hz and $\Omega_k(t)$ in cycles/octave are temporal and spectral ripple modulations, respectively, $x_i = \log_2[f_i/f_0]$ in octaves is a logarithmic frequency axis relative to $f_0$ in Hz, and $\Phi_{jk}$ and $\varphi_i$ are random initial phases.

For generating TORC-based stimuli, envelopes of seven “short” TORCs were first generated, each consisting of six ripples with temporal modulation: $\omega_j = 4j$ Hz (for $j = 1, \ldots, 6$), and each having a fixed ($k = 1$) spectral modulation: $\Omega = -1.5, -0.9, -0.3, 0, 0.6, 1.2$, and 1.8 cycles/octaves, respectively. All TORCs had rise and fall times of 5 msec, modulation depth of $\leq 30$ dB ($a_{jk} = 30/6$), and lasted for 250 msec. Such short TORC envelopes were then randomly adjoined to generate a “default envelope” that lasted for 4–5.5 sec.

To examine the effects of the changes in sound intensities in conditioning stimuli, we applied the default envelope to the carrier frequencies over a bandwidth of 5 octaves (0.88 to 28.16 kHz in steps of 1/128 octaves), scaled the peak amplitude to the speaker driver range, and then varied the amplitudes over the maximum range of 40 dB attenuation. To examine the effects of frequency changes, we generated a default envelope over the bandwidth of 2 or 4 octaves, and chose carrier sinusoidals within the range of 6 octaves (0.625 to 40 kHz in steps of 1/128 octaves, random phase at time zero) so that the signals had the same envelope with shifted bandwidth (e.g., 0.625–10 kHz, 2.5–40 kHz, and so on). We then normalized the peak amplitude of sound fragments with respect to their total signal powers ($\int |y(t)|^2 dt$) and uniformly scaled to fit them all within the speaker driver range.

**Dynamic Moving Ripples** To generate the DMR envelopes (Eq.3.2; $j = k = 1$), spectral modulations $\Omega(t)$ were sampled at 6 Hz from a uniform distribution in interval $\pm 1.5$ cycles/octave, and temporal modulations $\omega(t)$ were sampled at 3 Hz from a uniform distribution ranging between $\pm 25$ Hz, both of which were then up-sampled to 97.656 kHz using a cubic interpolation procedure (interp1 function with the “cubic” option in MATLAB).
Carrier frequencies were chosen as in the TORC stimuli, and applied to the envelope as in Eqs.(3.1)–(3.3). All DMR signals had rise and fall times of 5 msec and modulation depth of ≤ 30 dB. The peak amplitude was scaled in the same way as the TORC fragments (see above).

**Modulated Harmonic Tones**  The following equation was used to generate modulated harmonic tones:

\[ y_{\text{MHT}}(t) = A(t) \sum_{i=0}^{M-1} \cos\left[ \frac{2^i}{m} \phi(t) + \varphi_i \right], \]  

(3.4)

where \( A(t) \) is the envelope, and \( \varphi_i \) and \( \phi(t) \) are the initial and time-varying phases, respectively. In this study, \( M = 5 \) tones were combined with the density: \( m = 0.5 \) or 1 tone/octave. The derivative of the phase \( \phi(t) \) with respect to time gives the instantaneous frequency \( f(t) \):

\[ \frac{\partial \phi(t)}{\partial t} = 2\pi f(t). \]  

(3.5)

Normal distributions sampled at 48 Hz were used to generate the envelope \( A(t) \) and the instantaneous frequency \( f(t) \). The mean and standard deviation of \( A(t) \) were chosen to be [40–65] and [15–5] dB, respectively, whereas the mean \( f(t) \) ranged over 3 octaves (0.375–3 kHz) and the standard deviation from 1 to 1/3 octaves. We then up-sampled \( A(t) \) and \( f(t) \) to 97.656 kHz using a cubic interpolation procedure, and used Eqs.(3.4) and (3.5) to generate the signal \( y_{\text{MHT}}(t) \) with random initial phase \( \varphi_i \).

For examining the effects of amplitude changes, we generated a signal for given \( A(t) \) and \( f(t) \), normalized its peak amplitude within the speaker driver range, and varied the amplitudes over the maximum range of 60 dB attenuations. For examining the effects of frequency changes, we generated signals with a fixed \( A(t) \) but shifted \( f(t) \) by up to ±2 octaves, normalized their total signal powers, and uniformly scaled the signals to fit them all within the speaker driver range. For examining the effects of the changes in AM or FM, we used fixed mean \( A(t) \) and \( f(t) \). Before the up-sampling procedures, however, \( A(t) \) or \( f(t) \) were scaled to vary the modulation depth by up to 3-fold and/or limited the modulation rates to have the bandwidth of
4 Hz; i.e., the fastest and slowest modulation rates were [20–24] and [0–4] Hz, respectively. The synthetic signals were normalized with respect to their total power, and uniformly scaled to fit them all within the speaker driver range.

**Modulated Pink Noise** Starting from white noise $x_0(t)$, we used the following iterative procedures to produce modulated pink noise $y_{MPN}(t)$ that has asymptotically the same temporal and spectral modulation patterns as a target natural sound $y_{NS}(t)$. First, we computed the analytic signal of $y_{NS}(t)$ by using the Hilbert transform $\mathcal{H}[\cdot]$, and decomposed it into the envelope $A_{NS}(t)$ and the phase $\phi_{NS}(t)$:

$$y_{NS}(t) + \mathcal{H}[y_{NS}(t)] = A_{NS}(t) \exp[j\phi_{NS}(t)], \quad (3.6)$$

where $j^2 = -1$. Second, using the Fourier transform $\mathcal{F}[\cdot]$, we filtered the signal from the $(i - 1)$-th iteration $x_{i-1}(t)$ to have the same power spectrum as $y_{NS}(t)$:

$$\tilde{z}_i(\omega_k) = \tilde{x}_{i-1}(\omega_k) \cdot \left| \frac{\tilde{y}_{NS}(\omega_k)}{\tilde{x}_{i-1}(\omega_k)} \right|, \quad (3.7)$$

where $\tilde{y}(\omega) = \mathcal{F}[y(t)]$ denotes the signal $y$ in the Fourier domain. Third, we computed the analytic signal of $z_i(t) = \mathcal{F}^{-1}[\tilde{z}_i(\omega)]$ as in Eq.(3.6):

$$z_i(t) + \mathcal{H}[z_i(t)] = B_i(t) \exp[j\psi_i(t)], \quad (3.8)$$

where $B_i(t)$ and $\psi_i(t)$ are the envelope and the phase, respectively. Finally, we generated the signal for the $i$-th update as:

$$x_i(t) = A_{NS}(t) \cos[\psi_i(t)]. \quad (3.9)$$

That is, $x_i(t)$ is a pink noise with the envelope of the target $y_{NS}(t)$. In this study, we updated the synthetic signal 1000 times to generate modulated pink noise: $y_{MPN}(t) = x_{1000}(t)$. The signals
\( y_{NS}(t) \) and \( y_{MPN}(t) \) were normalized with respect to their total power, and then uniformly scaled to fit them all within the speaker driver range.

### 3.3 Data Analysis Methods

All data analysis was performed in MATLAB. As a preprocessing, we applied a median filter to clip spikes from the recording data (10 msec window), and centered the subthreshold responses to have zero mean (i.e., \( r(t) - \langle r(t) \rangle_t \), instead of subtracting the resting potential; \( \langle \cdot \rangle_t \) indicates the average over time \( t \); see also Figure 3.2).

The noise correlation was computed as the auto-correlation of additive noise \( \varepsilon_{ij}(t) \) (defined as in Eq.(3.12) and estimated as in Eq.(3.15); see also Eq.(A.85) on page 168):

\[
\varrho_{ij}(t) = \frac{\varepsilon_{ij}(-t) \ast \varepsilon_{ij}(t)}{\sigma_{\text{noise}}^2},
\]

where \( \ast \) indicates convolution. The correlation function \( \varrho_{ij}(t) \) represents a similarity in the noise signal over time, and thus a long correlation time suggests the existence of slow internal (stimulus-independent) dynamics in neural behaviors. However, our data showed that \( \varrho_{ij}(t) \) had a sharp peak with a rapid decay (within a few hundred milliseconds; Figure 3.2D) and thus a slow-time scale identified in neural responses in A1 (see Section 3.5 and Figure 3.11) would be governed by stimulus-dependent mechanisms. Together with the fact that the noise distribution agrees fairly well to the Gaussian distribution with the same standard deviation (Figure 3.2C), our data support the additive independent and identically-distributed (i.i.d.) Gaussian noise assumption in our model in Eq.(3.12).
Figure 3.2: **Recording data statistics.** (A) Distribution of recorded membrane potential (around its mean value). The mean membrane potential over time was first subtracted from entire recording data for each cell (i.e., \( r(t) - \langle r(t) \rangle_t \)), where \( \langle r(t) \rangle_t = -60.5 \pm 6.4 \) mV; mean ± standard deviation; 123 cells), and histogram was made in step of 0.1 mV; shown is then the mean probability distribution over the populations. The peak value (mode; \(-4.0 \) mV) would be a good estimate of the resting potential, i.e., \(-64.5 \) mV. (B) Distribution of mean correlation coefficients over trials (gray arrow head indicates the mean). All pairwise Pearson product-moment correlation coefficients were first computed between response traces to each stimulus fragment over trials, and the average was taken over pairs and stimuli for each cell; shown is the distribution of this average value over the population (0.27±0.09; mean ± standard deviation; 123 cells). (C) Distribution of additive noise components (estimated as in Eq.(3.15)). Although the noise distribution (black) has a slightly long tail in the positive direction (mode; \(-1.0 \) mV), it agrees fairly well to the Gaussian distribution (gray) with the same standard deviation (\( \hat{\sigma}_{\text{noise}} = 5.1 \) mV). (D) Noise correlation (Eq.(3.10)). Note the sharp peak at time zero with a rapid decay within a few hundred milliseconds. Together with (C), the results suggest that the additive i.i.d. Gaussian noise assumption (\( \varepsilon_{ij}(t) \sim \mathcal{N}[0, \sigma_{\text{noise}}^2] \)) more or less holds.
3.3.1 Context-Dependence

For those recordings that we could test at least four repeats on any given combinations of a probe stimulus and at least two conditioning stimuli, context-dependence—i.e., the response variability to a probe stimulus due to the presence of different conditioning stimuli—was examined in two ways; (1) significance measure in the statistics sense, and (2) fractional power measure in the response dynamics.

Significance Measure

For each sampled time point $t$ on a probe stimulus (with $t = 0$ indicating the transition from conditioning to probe stimuli), we performed a one-way non-parametric analysis of variance (Kruskal-Wallis test; Kruskal and Wallis, 1952) for equal medians among the responses $r_{ij}(t)$ over trials $j = 1,\ldots,m$ in all the contexts $i = 1,\ldots,n$ we examined; see also Eq.(3.12). Briefly, we first ranked the data $r_{ij}(t)$ into integers $r_{ij}(t)$ from 1 to $mn$ for each time point $t$. The test statistic $KW(t)$ is then given as:

$$KW = (mn - 1) \sum_i m \frac{\langle (r_{ij})_j - \langle (r_{ij})_j \rangle_i \rangle_j^2}{\sum_i \sum_j (r_{ij} - \langle (r_{ij})_j \rangle_i)^2} = \frac{3\langle (2\langle (r_{ij})_j - mn - 1 \rangle_j^2 \rangle_i}{mn + 1},$$

where $\langle \cdot \rangle$ indicates the average over trials with subscript $j$ and over contexts with subscript $i$, and the probability distribution of $KW$ can be approximated as a chi-square distribution with $n - 1$ degrees of freedom. (The time ($t$) is ignored here for brevity.)

Our criterion for the significance level was $p < 0.01$ for $\geq 5$ msec (i.e., at least 50 consecutive time points to avoid false positives due to multiple comparisons over time). Here we did not perform any post hoc analysis, partly because of data limitation, and partly because the goal of this analysis was to detect whether or not neurons showed context-dependence but not to identify what causes the difference in the response patterns—stimulus space is huge in general and thus we asked the latter question by taking a special stimulus design with synthetic sounds (see Sections 3.2.3–3.2.4); i.e., we fixed all the stimulus properties among condition-
ing stimuli but one so as to examine the effects of changes in a particular sound property of interest. The significance measure was then used in the population data analysis to compute the proportion—or, the probability—of observing significant context-dependence at a given moment. The noise floor—or, the level of false positive—was determined by resampling methods, where the trials were randomly shuffled to lose the information on the contexts, followed by the same significance test described above. For each probe, we repeated this procedure 1000 times, and took the average over the population to identify the chance level of declaring “significance” in this analysis.

**Fractional Power Measure**

A second measure was introduced to examine the contribution of context-dependence to response dynamics; i.e., a quantity meaningful from a modeling (instead of just statistical) perspective, based on the fraction of the response power that depends on stimulus history and its context (see also Section 3.3.3). We first assumed the following additive model for the response to a probe stimulus over time $t$:

$$ r_{ij}(t) = \mu(t) + \nu_i(t) + \varepsilon_{ij}(t). \tag{3.12} $$

That is, the observed response $r_{ij}(t)$ in the $i$-th context for the $j$-th trial consists of i.i.d. Gaussian noise: $\varepsilon_{ij}(t) \sim \mathcal{N}[0, \sigma_{\text{noise}}^2]$—with zero mean and the variance of $\sigma_{\text{noise}}^2$—and stimulus-related (predictable) parts, which can be further decomposed into context-dependent and independent fractions: $\nu_i(t)$ and $\mu(t)$, respectively. Here, each component can be estimated as:

$$ \hat{\mu}(t) = \langle \langle r_{ij}(t) \rangle_j \rangle_i \tag{3.13} $$

$$ \hat{\nu}_i(t) = \langle r_{ij}(t) \rangle_j - \langle \langle r_{ij}(t) \rangle_j \rangle_i \tag{3.14} $$

$$ \hat{\varepsilon}_{ij}(t) = r_{ij}(t) - \langle r_{ij}(t) \rangle_j \tag{3.15} $$
where \( \langle \cdot \rangle \) indicates the average over probes (with no subscript).

Then the measure was defined as the context-dependent fractional power: \( P[\nu_i(t)] \equiv \langle \langle \nu_i^2(t) \rangle \rangle \), normalized by the predictable response power: \( P[\mu(t) + \nu_i(t)] \). In the infinite data limit (i.e., every possible probe examined with all contexts), the predictable response power would become time-invariant: \( P[\mu(t) + \nu_i(t)] \rightarrow \sigma^2 \equiv \langle P[\mu(t) + \nu_i(t)] \rangle_t \), where the symbol "\( \equiv \)" means “equal in expectation.” In practice, however, the population average of the response power was nonstationary over time—typically, with large fluctuation soon after the transition from conditioning to probe stimuli (Figure 3.5B) because of a finite recording length and stimulus design—and thus the predictable response power (the denominator of Eqs.(3.21) and (3.23) shown below) was smoothed by taking the running average over \([0, 2t]\) at time \( t \). Note that the “power” is usually computed as the average over time, but we here assumed ergodicity and thus took the average over the populations.

The fractional response powers can be estimated by considering the “average power” and the “power of the average”, under the assumption that the additive components in Eq.(3.12) are all uncorrelated between each other at any given moment (see also Sahani and Linden, 2003). Considering the average over trials (for \( j = 1, \ldots, m \)), we have:

\[
\langle \langle \nu_i^2(t) \rangle \rangle \equiv P[\mu(t) + \nu_i(t)] + \langle \langle \varepsilon_{ij}^2(t) \rangle \rangle_i,
\]

\[
\langle \langle r_{ij}(t) \rangle \rangle^2_i \equiv P[\mu(t) + \nu_i(t)] + \langle \langle \varepsilon_{ij}(t) \rangle \rangle^2_i,
\]

(3.16)

(3.17)

From central limit theorem, we have: \( \langle \langle \varepsilon_{ij}(t) \rangle \rangle^2_i \equiv \langle \langle \varepsilon_{ij}^2(t) \rangle \rangle_i/m \), and thus:

\[
\hat{P}[\mu(t) + \nu_i(t)] = \left\langle \left\langle \frac{m \langle r_{ij}(t) \rangle^2_j - \langle r_{ij}^2(t) \rangle_j}{m - 1} \right\rangle \right\rangle_i,
\]

(3.18)

where \( \langle \cdot \rangle \) (without subscripts) indicates the average over all the tested probe stimuli in the population data. Similarly, considering the average of the trial average \( \langle \langle r_{ij}(t) \rangle \rangle \) over contexts
$(i = 1, \ldots, n)$, we have:

$$
\hat{P}[\mu(t)] = \left\langle \frac{n\langle r_{ij}(t)\rangle_{j}^{2} - \langle r_{ij}(t)\rangle_{j}^{2}}{n - 1} \right\rangle_i.
$$

(3.19)

Therefore, from Eqs.(3.18) and (3.19), the context-dependent fractional power can be given in expectation as:

$$
\hat{P}[\nu_i(t)] = \hat{P}[\mu(t) + \nu_i(t)] - \hat{P}[\mu(t)],
$$

(3.20)

and we could use the following quantity as a measure of the contribution of context-dependence to response dynamics:

$$
\frac{\hat{P}[\nu_i(t)]}{\hat{P}[\mu(t) + \nu_i(t)]}.
$$

(3.21)

The denominator was smoothed by taking the moving average over $[0, 2t]$ at time $t$ before computing Eq.(3.21) (and Eq.(3.23) in Section 3.3.3).

This fractional power measure differs from the significance measure in two ways. First, the former is continuous over time, whereas the latter involves arbitrary thresholding procedure to determine the significance level and is consequently binary. Second, the fractional power measure involves the normalization (by the stimulus-related power; $\mathcal{P}[\mu(t) + \nu_i(t)]$) and thus depends on the “relative” difference to the overall fluctuation between the response patterns caused by stimulus history, whereas the significance measure (and the numerator of the fractional power measure; i.e., context-dependent fraction; $\mathcal{P}[\nu_i(t)]$) depends on the “absolute” differences. Therefore the fractional power measure can be said to be more faithful in these respects (see also the relation to the response predictability in Section 3.3.3).
3.3.2 Exponential Curve Fit

To measure the relevant time-scales of the context-dependence, we fit a (sum of) exponential processes to the population data (Figures 3.5, 3.8, and 3.11):

\[ \alpha + \sum_k \alpha_k \exp \left( -\frac{t}{\tau_k} \right), \]

where \( \alpha_k \) and \( \tau_k \) indicate the decay size and constant, respectively. We used the lsqcurvefit function in the MATLAB Optimization Toolbox for the curve fitting, and used the following criterion for choosing the number of exponential processes: \( |\alpha_k| > \sum_k |\alpha_k| / 10 \) for all \( k \), that is, the contribution of an exponential process must be at least one-tenth of the total. Confidence intervals of the parameters were computed by resampling methods (200 repeats with randomly selected 1000 samples; Figure 3.9 on page 118).

3.3.3 Response Predictability

To analyze how response predictability in A1 depends on stimulus history and its context over time, we computed the time-course of the ratio between context-independent fractional power, \( \mathcal{P}[\mu(t)] \), and the stimulus-related response power, \( \mathcal{P}[\mu(t) + \nu_i(t)] \). From a modeling viewpoint, the former is the best we could reach in the response estimation exploiting the stimulus history for a limited duration of \( t \), whereas the latter is the very best level that no model could outperform under the additive noise assumption. (Note that the context-dependent response power \( \mathcal{P}[\nu_i(t)] \) is the fraction that is not accessible when only a finite stimulus history is available, and that the noise power \( \sigma^2_{\text{noise}} = \langle \mathcal{P}[\varepsilon_{ij}(t)] \rangle_t \) is the fraction that is never accessible under the additive noise assumption.) Therefore, the following ratio indicates the context-dependence of the response predictability, which constitutes an upper-bound estimate of the response prediction
performance given a window length \( t \):

\[
\frac{\hat{P}[\mu(t)]}{\hat{P}[\mu(t) + \nu(t)]}.
\]

This Eq.(3.23) was computed from Eqs.(3.18) and (3.19).

### 3.4 Encoding Models

Several neural encoding models were examined in this study to compare their performance to the upper-bound estimate (Eq.(3.23)) for further analyzing the context-dependence of the response predictability. Specifically, we used STRF-based models (Section 3.4.1) as well as general high-order models such as artificial neural networks and support vector regression (Section 3.4.2). Such generic models do have some problems in their interpretations from biological viewpoints, but we exploited these models simply for examining the data fit performance. Here we (re)analyzed the recording data from the previous work (20 cells; Machens et al., 2004) because the recording data we collected for examining context-dependence were not tested with enough varieties of stimuli (Sections 3.2.3–3.2.4), which could cause a bias in the parameter estimation (Paninski, 2003a; Simoncelli et al., 2004; also see Figure 3.10); the total length of distinct stimulus fragments of this and the previous works was typically \( \sim 40 \) sec and \( > 5 \) min, respectively.

#### 3.4.1 Linear-Nonlinear Cascade Models

Spectro-temporal receptive field (STRF; Klein et al., 2000) models can be formulated in general as:

\[
\hat{r}(t) = \iint \text{STRF}(\tau, \omega) \cdot S(t - \tau, \omega) \, d\tau \, d\omega,
\]  

(3.24)
where $\hat{r}(t)$ is the estimated response, and $S(t, \omega)$ is the spectrogram (short-time Fourier transform; Cohen, 1995) of the sound pressure waveform $s(t)$:

$$S(t, \omega) = 20 \log_{10} \left| \frac{1}{\sqrt{2\pi}} \int e^{-j\omega \tau} s(\tau) h(\tau - t) d\tau \right|.$$  \hspace{1cm} (3.25)

The energy of a window function $h(t)$ is typically taken to be unity ($\int |h(t)|^2 dt = 1$) to ensure that the energy of the spectrogram $S(t, \omega)$ is equal to that of the signal $s(t)$.

For the parameter estimation, we exploited linear regression techniques (for details, see Appendix Section A.2). Specifically, we first discretized time $t$ and frequency $\omega$, and re-ordered the indices to simplify Eq.(3.24) into the following form:

$$\hat{r} = S\beta,$$ \hspace{1cm} (3.26)

where $\hat{r}$ and $\beta$ are column vectors of the estimated response and STRF, respectively, and the $i$-th row of the matrix $S$ consists of the $i$-th stimulus vector. Ridge regression was then used to estimate the STRF (parameters): $\beta$ (for details, see Appendix Section A.2.1).

In this study we performed 10-fold cross-validation, i.e., split the data set into training (90%) and validation (10%) data sets, used the training data set to estimate STRF with various power constraint values ($\lambda$ in Eq.(A.53) on page 154), and chose the one ($\hat{\beta}_{\text{ridge}}$) that gave the best model performance on the validation data set (see below Eq.(3.29) in Section 3.4.3). The resulting model performance on the training and validation data set can then be considered as the upper and lower estimates, respectively.

To vary the window length while fixing the model complexity—i.e., the number of free parameters in a model—for a fair comparison of the model performance (Figure 3.11), the time bin sizes were varied in a pseudo-logarithmic scale: $\Delta_k t = 2^k$ msec for $k = 2, \ldots, 10$ from near to distant past. In this study we set the number of bins for $\Delta_k t$ (45 in total) as $[45, 0, \ldots, 0], [9, 36, \ldots, 0], \ldots, [9, 8, \ldots, 1]$, resulting in models with window lengths of 180, 324,
548, 884, 1364, 2004, 2772, 3540, and 4052 msec, respectively. Frequency discretization was \( \Delta x = 3 \) bins/octave, ranging between 0.4–22 kHz (18 frequency bins).

**Static Nonlinearity**

Static nonlinearities can be given as a nonlinear transformation \( g_{sn} \) that acts on the output of the linear model (Eq.(3.24)) to form a new (better) estimate (see e.g., Simoncelli et al., 2004; Machens et al., 2004):

\[
g_{sn} : \hat{r}(t) \mapsto \hat{r}_{sn}(t). \tag{3.27}
\]

Here we used a scatter plot smoothing technique to identify the best transformation \( \hat{g}_{sn} \) on the training data set (for details, see Appendix Section A.3.1)—resulting in the upper estimate of the model performance—and then applied it to the validation data set for computing the lower estimate.

**3.4.2 Nonlinear Models**

In addition to classical STRF-based models, the performance of two generic nonlinear models was also examined; artificial neural networks (multi-layer perceptrons) and support vector regression (for details, see Appendix Section A.3). These models typically have high expressive power—in fact, they can provide asymptotically optimal solutions to an arbitrary fitting problem under appropriate conditions (Hecht-Nielsen, 1989; Hertz et al., 1991)—and thus we asked if the model performance was comparable to the upper bound estimate of the response predictability with a given window length. As before, here we used 10-fold cross-validation to find appropriate parameters, although it was not very successful for support vector regression (Table 3.2). Because of a limited data length and the huge computational power (and memory resources) required for the training, we only examined models with window lengths of up to several hundred milliseconds.
Multi-Layer Neural Networks

Here we used classical three-layer (input, hidden, and output) feedforward networks, with log-sigmoid and linear transfer functions for hidden and output units, respectively (see Appendix Section A.3.2). The time discretization was $\Delta t = 4$ msec, and we tested the model with window lengths of 40, 200, and 500 msec for spectrogram-based models with a frequency discretization of $\Delta f = 3$ bins/octave ranging from 0.4 to 22 kHz (18 frequency bins). We also assessed cochleagram-based models (window length: 200 msec) with 100 frequency bins (ranging between 0.4–22 kHz) in the equivalent rectangular bandwidth (ERB) space (Moore and Glasberg, 1983). Here we used a Gammatone filter bank to generate the cochleagrams of input sounds (Auditory Toolbox for MATLAB; Slaney, 1993).

As a preprocessing, we performed principal component analysis (see Appendix Section A.1.1) to whiten the data, and chose those principal components that capture at least 0.1% of total variance (or at least 100 principal components in total) for further processing.

For training the neural networks, we started with 12 hidden units and a single output unit, and trained the network with a conjugate-gradient backpropagation algorithm (MATLAB Neural Network Toolbox; see also Appendix Section A.3.2). Then the least contributing unit was pruned one by one unless the performance on the validation data set got worse, typically ending up with nine or ten hidden units. The training procedures were restarted with 10 random initial conditions, and the one with the best performance (see Section 3.4.3) was chosen as a final result.

Support Vector Regression

As an alternative to the multi-layer perceptrons, we examined the performance of support vector regression ($\nu$-SVR; Schölkopf et al., 2000) with gaussian kernels, which in fact is equivalent to radial basis function networks (see also Appendix Section A.3.3). An open-source library was used for the training (LIBSVM; Chang and Lin, 2001).
As before, the time and frequency discretizations were $\Delta t = 4$ msec and $\Delta x = 3$ bins/octave (ranging from 0.4 to 22 kHz), and the model had a window length of 200 msec. Before training the model, the data were scaled to have zero mean and unit variance.

To find appropriate parameters, a coarse grid search ($1/\lambda$ and $\gamma$ for $10^{-4}$, $10^{-2}$, $10^0$, and $10^2$ (four steps in $[-4, 2]$ in log-space); see also Appendix Section A.3.3) was first performed with a random subset of 5000 data for each cell, and then a fine grid search (three steps in $[-1, 1]$ in log-space, centered around the best parameter value from the coarse grid search) with another random subset of 5000 data. Such a two-step grid search was performed for $\nu \in [0.1, 0.5]$ in steps of 0.1, and the best parameter set obtained ($\lambda$, $\gamma$ and $\nu$) was used for training a $\nu$-SVR with the entire training data set. However, support vector regression tends to overfit a training data set as far as we tested (see Table 3.2) partly because of a limited data length and partly because of inappropriate choices of the parameters (and kernels/models).

### 3.4.3 Model Performance

The performance of the neural encoding models described above was quantified as the ratio between the estimated response power captured by a model, $\sigma^2_{\text{model}}$, and the stimulus-related (predictable) response power, $\sigma^2$ (Sahani and Linden, 2003; Machens et al., 2004). Note the similarity to the analysis of response predictability in Section 3.3.3.

Assuming additive i.i.d. Gaussian noise: $\varepsilon_j(t) \sim \mathcal{N}[0, \sigma_{\text{noise}}^2]$ over trials (for $j = 1, \ldots, m$) and time $t$, we can express the observed response for the $j$-th trial as: $r_j(t) = \rho(t) + \varepsilon_j(t)$, with the stimulus-related components $\rho(t)$ (equivalent to $\mu(t) + \nu(t)$ in Eq.(3.12)), and the total power in the observed response as: $\sigma^2_{\text{total}} = \sigma^2 + \sigma_{\text{noise}}^2$, with the stimulus-related power: $\sigma^2 \overset{\text{def}}{=} \langle \rho^2(t) \rangle_t$ in the limit of large $t$. (As before, we use $\langle \cdot \rangle_t$ to indicate the average over time with subscript $t$ and over trials with subscript $j$. Note that $\langle \rho(t) \rangle_t = 0$, because of the preprocessing of $r_j(t)$ to have zero mean.)
From central limit theorem, the power of the average response over trials can be written as 
\[ \langle \langle r_j(t) \rangle \rangle_t^2 = \sigma^2 + \sigma^2_{\text{noise}}/m. \] Hence, the predictable response power \( \sigma^2 \) can be estimated as:
\[ \hat{\sigma}^2 = \left\langle \frac{m\langle r_j(t) \rangle_t^2 - \langle r_j^2(t) \rangle_j}{m - 1} \right\rangle_t, \] (3.28)
where we use \( \hat{\sigma}^2_{\text{total}} = \left\langle \langle r_j^2(t) \rangle \rangle_t \) and note the similarity to Eqs.(3.18) and (3.19).

The model performance \( \sigma^2_{\text{model}}/\sigma^2 \) is then given as:
\[ \frac{\hat{\sigma}^2_{\text{total}} - \hat{\sigma}^2_{\text{error}}}{\hat{\sigma}^2}, \] (3.29)
where \( \hat{\sigma}^2_{\text{error}} = \left\langle \langle (r_i(t) - \hat{r}_i(t))^2 \rangle \rangle_t \) is the model error power. In Eq.(3.23) the average was taken over the population but not over the time, but the quantities in Eqs.(3.23) and (3.29) are equivalent under the assumption of ergodicity.

### 3.5 Results

To characterize the time course and magnitude of context-dependent effects on neural responses in rat primary auditory cortex (area A1), neurons were probed with a variety of spectro-temporally rich stimuli (e.g., animal vocalizations) in sequence (see Section 3.2 and also Figure 3.1). The use of such complex stimuli allowed us to probe a larger fraction of stimulus space than conventional protocols using tones and others (e.g., Theunissen et al., 2000, 2001; Sen et al., 2001; Bar-Yosef et al., 2002; Machens et al., 2004; Garcia-Lazaro et al., 2006).

Our analysis consisted of the following three parts. First, we assessed the overall context-dependence of neurons in A1 using natural sound ensembles (Section 3.5.1). Second, we used synthetic sounds to characterize how context-dependence depended on stimulus properties such as stimulus intensity and modulation rates (Section 3.5.2). Finally, we further quantified the context-dependence from the viewpoint of model construction, i.e., measured the response predictability given all the past stimulus information within an arbitrary window.
length (Section 3.5.3). Note that here we only estimated the upper-bound of the prediction performance, and do not have an actual encoding model yet that can perfectly predict neural responses in A1 from acoustic signals (but see approaches in Section 3.5.4).

### 3.5.1 Context-Dependence

Firing rates in A1 were typically low under our experimental conditions (Figure 2.12 on page 56; spontaneous, $0.25\pm0.65$ Hz; evoked, $0.33\pm0.69$ Hz; median ± interquartile range for 194 cells; see also Wehr and Zador, 2005). We therefore examined subthreshold responses rather than firing rates—because subthreshold responses consist of a continuous variable in time (membrane potential) rather than a sparse binary time series (a train of action potentials), we could obtain good estimates of activity even in the complete absence of spiking activity. Note however that we did not analyze in detail the membrane potential dynamics at “silence” periods because spontaneous activity prevented us from detecting especially the long-lasting events saliently, even though a subset of cells showed significant context-dependent fluctuations soon after the onset of inter-sequence intervals (Figure 3.3).

Figure 3.4 shows a typical example of subthreshold responses to a six-second natural sound stimulus in two different natural sound contexts, i.e., preceded by two different six-second conditioning stimuli. Consistent with previous work (Machens et al., 2004), this neuron showed high trial-to-trial reliability (Figure 3.4A) within each set of trials for which the conditioning stimuli were held fixed—the correlation coefficient of the response traces across trials in a given context was $0.61\pm0.07$ (mean ± standard deviation) for the seven natural sound fragments tested in this cell. The reliability varied within a given neuron as a function of the stimuli tested, and across neurons; the mean correlation coefficient was $0.31\pm0.09$ (mean ± standard deviation; 39 cells) over the population examined with natural sound ensembles.

Changing the conditioning stimulus—i.e., the stimulus context—caused a dramatic change in the response to the probe stimulus (Figure 3.4B, blue vs. red). In this example, the effects of the context on the response lasted more than four seconds. Interestingly, the
context-induced differences could sometimes be intermittent; in this example, the two average response traces showed no difference in the interval two to four seconds after the onset of the probe, but diverged again at around four seconds.

Here we used two measures to quantify the differences in the probe stimulus induced by context. The first (Figure 3.4D) assessed whether the differences in the traces were statistically significant ($p < 0.01$ for at least 5 msec; see Section 3.3.1), whereas the second method (Figure 3.4E) assessed the component of the response power (variance at a given time) dependent on stimulus history (Eqs.(3.18)–(3.21) in Section 3.3.1). The two measures generally agreed quite well, as can be confirmed by noting that when the two traces were significantly different (vertical gray strips), the power was typically high.

Although context-dependent effects could manifest intermittently in a given cell (as in Figure 3.4), across the population these effects showed an orderly monotonic decay (Figure 3.5). Of 305 natural sound probe stimuli tested with different—typically, around five to eight—natural sound contexts in 39 cells, significant effects were observed in 204 cases (66.9%; Figure 3.5A), and about a third (30.3%) of the events occurred longer than one second after the onset of a probe stimulus (Figure 3.5A). This fraction represents a lower bound on the maximum duration of the possible effect in a given cell, since the number of conditioning-probe combinations tested per neuron was quite small, and was not tailored to the properties of the cell. For both measures there was a long decay constant of about one second ($\tau = 0.90$ and 1.04 sec, respectively; see Figure 3.5 for details). Since this decay represents the dynamics of the average response—we did not have enough data per neuron to reliably estimate the time constant for each neuron individually—it could include some responses with longer constants (e.g., Figure 3.4) and some with shorter. We found no systematic relationship between the duration or strength of context-dependence and the recording depth or cortical layers (Figure 3.6; see also Ulanovsky et al., 2004). In conclusion, a neuron’s “memory” of the stimulus history it has experienced can persist for at least one or several seconds.
Figure 3.3: **Context-dependent effects on spontaneous dynamics.** A subset of cells in A1—typically those with low spontaneous activities as in this example—showed context-dependent fluctuations soon after the onset of inter-sequence intervals (time zero indicates the termination of stimuli or the onset of silence periods). Spikes were clipped by a median filter (window length: 10 msec). In this example, the spontaneous dynamics—examined at a silence period preceded by three different stimuli (blue, red, and green; median over 12 trials each)—diverged several hundred milliseconds after the onset of the silence (black line, \( p \)-value for Kruskal-Wallis test; gray vertical strip indicates time points where \( p < 0.01 \) for \( \geq 5 \) msec). For details, see Section 3.3.1 and also Figure 3.4.
Figure 3.4: **Context-dependence of neural responses lasted for several seconds.** (A) Typical subthreshold responses of a rat A1 neuron to part of a natural sound sequence (*spectrogram*; time zero indicates the transition from *conditioning* to *probe* stimuli) over six repeats (*red* lines). Spikes were clipped by a median filter (window length: 10 msec). The neuron showed high trial-to-trial reliability (correlation coefficients across trials to this particular probe stimulus, $0.74 \pm 0.08$; and across trials across all natural sound stimuli examined in this cell, $0.61 \pm 0.07$; mean $\pm$ standard deviation). (B) The median responses to the probe stimulus in two different contexts; *red* line for the one shown in (A), and *blue* line for the one in response to the same probe stimulus but preceded by a “silence” period in this example. Note however that silence is not always the “exception” that causes a difference in the following response dynamics. Significant dependence on the stimulus history was observed for longer than four seconds (*gray* bands; see (D) for details), whereas the responses between two to four seconds after the onset of the probe stimulus were not so much dependent on the conditioning stimuli we tested. (C) The recorded cell in this example was histologically identified as a layer II-III pyramidal neuron (scale-bar; 100 $\mu$m). (D) We performed a pointwise statistical (Kruskal-Wallis) test for equal medians between the responses to the probe stimulus in all different contexts (*black* line; $p$-values), and the *gray* bands show the time points where the context-dependence was statistically significant under the criterion: $p < 0.01$ for $\geq 5$ msec. (E) The response power that depends on the context (*black* line; $\langle \hat{\nu}_i^2(t) \rangle_i$ from Eqs.(3.18)–(3.20) without population average; see Section 3.3 for details) well represents the magnitude of the context-dependence. The population average of this quantity with and without normalization by the average predictable power is shown in Figure 3.5B.
Figure 3.5: Context-dependence can last for seconds. (A) Significance measure. Top: Each raster shows periods during which the significance measure exceeds threshold ($p < 0.01$ for $\geq 5$ msec; see Section 3.3) for a particular probe-stimulus combination in a given neuron (an example is shown as gray bands in Figure 3.4). The rasters are sorted according to the longest-lasting effect, so successive rasters may correspond to different neurons. Significant context-dependence was observed in about two-thirds of probe stimuli (204 out of 305 sets in 39 neurons). Bottom: The black curve shows the proportion—or the probability—of observing the significant context dependence, and the orange curve shows the noise floor computed by resampling methods. The former is well fit by the sum of two exponentials (gray: $\alpha_1 = 0.17$, $\tau_1 = 0.20$ sec, $\alpha_2 = 0.09$, and $\tau_2 = 0.90$ sec for Eq.(3.22) with the mean noise floor over time $\alpha = 2.8 \times 10^{-3}$). Around 30% of the context-dependent events occurred at $\geq 1$ sec from the onset of probe stimuli (brown; cumulative probability). (B) Fractional power measure. Top: From the same population data, we computed the stimulus-related response power (black, $\hat{P}[\mu(t) + \nu_i(t)]$ in Eq.(3.18); gray, its moving average over the data in $[0, 2t]$ at time $t$) and the fraction that depends on stimulus history and its context (thick; $\hat{P}[\nu_i(t)]$ in Eq.(3.20)). Note that the latter corresponds well to the significance measure as shown in the bottom panel of (A). See also Figure 3.1A. Bottom: The ratio of the context-dependent power to the stimulus-related power represents well the contribution of stimulus history to the response dynamics (black; Eq.(3.21)). The decay size and constant are: $\alpha_1 = 0.49$, $\tau_1 = 1.04$ sec, with $\alpha = 0$ (gray). The absence of a fast time-scale suggests that many of the fast components in (A) resulted from the large fluctuation in the response dynamics induced by the transition from conditioning to probe stimuli, while slow components were by and large left intact.
Figure 3.6: **No clear correlation between recording depth and longest-lasting context-dependence effects.** Each dot represents the longest-lasting effect examined with natural sound conditioning-probe combinations for each recorded cell, as a function of recording depth as determined from micro-manipulator travel (39 cells), and the histograms are shown in *blue* lines. There was no clear relation between the two quantities, consistent with previous work (see also Ulanovsky et al., 2004)
3.5.2 Relation to Acoustic Properties

We initially chose natural sounds because of their spectro-temporally rich melange of acoustic structures, as we expected that the changes in such a full variety of sound properties would allow us to observe as large and long-lasting effects as possible. However, a problem is that we could not tell by comparing two or more conditioning natural sound fragments what exactly made neurons in A1 behave differently in response to the same probe stimulus.

To study what stimulus properties contribute to the context-dependence, we then used synthetic stimuli for which we can control the stimulus parameters well. To examine the frequency change effects among conditioning stimuli for example (Figure 3.7), we first generated a fragment of synthetic sounds—e.g., dynamic moving ripples (Eqs.(3.1)–(3.3) on page 90)—and then up- or down-shifted its spectral components over the maximum range of 4 octaves to make (conditioning) stimulus ensembles where only a particular sound property of interest—frequency in this example (Figure 3.7A)—is different but all the other parameters—e.g., stimulus intensities and modulation patterns—are the same. In this and similar ways, here we examined the effects of the changes in each of the following acoustic properties: amplitude (over a maximum range of 60 dB attenuation), frequency (with a maximum shift of 4 octaves), amplitude-modulation (AM; with a maximum difference in modulation rate and depth of 20 Hz and 3-fold, respectively), frequency-modulation (FM; with the same maximum difference range as for AM), and higher-order spectro-temporal structures (by comparing natural sounds and corresponding modulated pink noise; Eqs.(3.6)–(3.9) on page 93); for details, see Sections 3.2.3–3.2.4, Table 3.1, and Figure 3.1B.

Figures 3.8 and 3.9 show the population results and the summary of the analysis, respectively. When we varied either intensities or frequencies (i.e., lower-order sound properties) in conditioning stimuli, the context-dependent effects were observed in 77 out of 95 cases (81.1%; 31 cells) and in 73/110 (64.6%; 35 cells), respectively (Figure 3.8A). The effects were almost as large and long as those caused by the difference between natural sound contexts (Figure 3.9B)—i.e., by the changes in any possible acoustic feature. However, the slow
components in the significance measure were less apparent in the context-dependence caused by the changes in the lower-order sound properties (Figure 3.9A). Single exponentials with the time-scale of several hundred milliseconds explained the effects in these two cases well (Figure 3.8A).

We then examined the effects of AM and FM changes in conditioning stimuli using modulated harmonic tones (Eqs.(3.4) and (3.5) on page 92), and also the changes in even higher-order acoustic properties such as complex interactions between spectro-temporal components by comparing the differences between modulated pink noise and its corresponding natural sounds (see Methods Section 3.2.4 and Figure 3.1B). The context-dependent effects were observed in 62 out of 96 cases (64.6%; 27 cells), in 47/82 (57.3%; 25 cells), and in 59/137 (43.1%; 27 cells), respectively, but the effects were substantially smaller and shorter than the effects caused by the changes in natural sound contexts in both measures (Figure 3.9); i.e., higher-order sound properties contributed to the context-dependence mainly on shorter time-scales (on the order of 100 msec). We thus conclude from Figures 3.8 and 3.9 that neural behaviors in A1 are more susceptible to changes in lower-order sound properties such as overall intensities and frequencies than to changes in higher-order properties including AM and FM.

3.5.3 Relation to Response Predictability

Population data showed that a time-scale relevant for A1 neurons was as slow as on the order of seconds (Figure 3.5). But, how important would past events be for predicting neural responses? To address this question, we estimated the best possible performance for the response prediction when only a finite length of stimulus history is available.

We first assumed that the observed responses to a given probe stimulus consisted of additive noise and stimulus-related components that can be further decomposed into context-dependent and independent fractions (Eq.(3.12) on page 97). We could then estimate the stimulus-related—or the predictable—fraction as the trial average in each context, whereas
the context-independent fraction as the mean response over all contexts (for details, see Section 3.3.3). The latter gives the best estimate of the responses to a probe stimulus without any knowledge on the conditioning stimuli, whereas the former corresponds to the very best response estimate that no model could outperform under the additive noise assumption. Because the prediction performance can be assessed by examining how much of the predictable response power a given model can capture (Linden et al., 2003; Sahani and Linden, 2003; Machens et al., 2004), we could then estimate the upper-bound of the prediction performance (given a finite window length) as the ratio of the response powers between the context-independent and the predictable fractions (Eq.(3.23) on page 101). Note the relation to the fractional response power measure for context-dependence, where we computed the ratio of the response powers between the context-dependent and stimulus-related fractions (Figures 3.5B and 3.8B; Eq.(3.21) on page 99).

The estimated upper-bound (blue curve in Figure 3.11) then indicates that no model could capture more than a half \((1 - |\alpha_1| = 0.51)\) of the response power given the window length of \(<100\) msec, and that the stimulus history for at least several seconds should be considered to predict neural responses in A1 fully enough \((\tau = 1.04\) sec). Such long-lasting effects would not be simply explained—or “expected”—by some “trivial” internal processes of neural dynamics because the noise correlation (Eq.(3.10) on page 94) had a sharp peak with a rapid decay (within a few hundred milliseconds; Figure 3.2D). Stimulus-dependent bottom-up mechanisms would thus be involved in the modulation of neural responses in A1 over seconds (see also Section 3.6.3).
Figure 3.7: **Context-dependence caused by frequency changes.** (A) Typical median sub-threshold responses of a rat A1 neuron to part of a sound sequence, where a natural sound stimulus was probed with two different synthetic conditioning stimuli with different bandwidths (*blue* line for a context with 0.625–2.5 kHz (*top spectrogram*), and *red* line for another context with 10–40 kHz (*bottom spectrogram*); time zero indicates the transition from conditioning to probe stimuli). Spikes were clipped by a median filter (window length: 10 msec). Although the responses were less reliable in this example (correlation coefficients across trials to this particular probe stimulus, 0.28±0.06; and across trials across all stimuli examined in this cell, 0.21±0.11; mean ± standard deviation) than those in Figure 3.4, significant context-dependence was observed for longer than four seconds as well (*gray* bands; see (B) for details). (B) *Black* line represents *p*-values and *gray* bands show the time points where the context-dependence was statistically significant under the criterion: *p* < 0.01 for ≥ 5 msec (in the same format as Figure 3.4D). (C) The response power that depends on the context (*black* line; in the same format as Figure 3.4E). The population average of this quantity normalized by the average predictable power is shown in Figure 3.8B (*frequency*).
Figure 3.8: **Relation between context-dependence and sound properties.** Using synthetic sounds, we examined the effects of the changes in a particular sound property of interest on the responses to following probe stimuli. From top to bottom panels, shown are the population analysis for the changes in amplitude (95 sets in 31 cells; red), frequency (110 sets in 35 cells; blue), AM rates and depths (96 sets in 27 cells; magenta), FM rates and depths (82 sets in 25 cells; cyan), and higher-order properties (137 sets in 27 cells; green). The black curves show the exponential curve fit as in Eq.(3.22) with the constant $\alpha$ being the mean noise floor over time and zero for (A) and (B), respectively. **(A) Significance measure.** Shown is the population analysis based on the probability of observing context-dependence, in the same format as Figure 3.5A but overlaid. The orange curves show the noise floor, and from top to bottom panels, the parameters for the exponential model (black) were $[\alpha_1, \tau_1, \alpha] = [0.23, 0.25, 2.8 \times 10^{-3}]$, $[0.19, 0.27, 2.3 \times 10^{-3}]$, $[0.13, 0.17, 2.3 \times 10^{-3}]$, $[0.07, 0.15, 2.1 \times 10^{-3}]$, and $[0.04, 0.13, 3.6 \times 10^{-3}]$, respectively. **(B) Fractional power measure.** Shown is the population analysis, indicating the contribution of context-dependence to the response dynamics, in the same format as Figure 3.5B. From top to bottom panels, the parameters for the exponential curves (black; all with $\alpha = 0$) were $[\alpha_1, \tau_1] = [0.57, 0.66]$, $[0.54, 0.97]$, $[0.86, 0.11]$, $[0.64, 0.15]$, and $[0.52, 0.20]$, respectively.
Figure 3.9: **Population analysis summary.** Parameters for the exponential curves (Eq.(3.22); top, $\sum_k \alpha_k$; bottom, $\tau_k$) in Figures 3.5 and 3.8 are shown with 95% confidence intervals computed by resampling methods (see Methods Section 3.3.1).  

(A) **Significance measure.** From left to right, the intervals are; $\sum_k \alpha_k \in [0.23, 0.28], [0.21, 0.25], [0.17, 0.22], [0.11, 0.14], [0.06, 0.09], and [0.03, 0.06]$, respectively; and $\tau_k \in \{[0.17, 0.24] and [0.74, 1.02]\}, [0.23, 0.28], [0.24, 0.30], [0.16, 0.18], [0.13, 0.16], and [0.09, 0.23]$, respectively.  

(B) **Fractional power measure.** From left to right, the intervals are; $\sum_k \alpha_k \in [0.46, 0.52], [0.52, 0.63], [0.51, 0.57], [0.82, 0.90], [0.58, 0.70], and [0.45, 0.61]$, respectively; and $\tau_k \in [0.91, 1.18], [0.54, 0.81], [0.85, 1.09], [0.10, 0.12], [0.12, 0.18], and [0.16, 0.25]$, respectively. Abbreviations; Nat, Natural Sounds; Amp, Amplitude; Freq, Frequency; AM, amplitude-modulation; FM, frequency-modulation; High, higher-order property.
3.5.4 Model Performance

This long time-scale explains in part why classical linear encoding (spectro-temporal receptive field; STRF) models with a limited window length, typically a few hundred milliseconds, have failed. The performance of STRF-based models was in general unsatisfactory, consistent with previous work (∼20%; Sahani and Linden, 2003; Machens et al., 2004). The performance did not improve significantly, however, just by extending the window length (up to ∼4 sec; red bands in Figure 3.11 for mean lower and upper bound estimates) even with static nonlinearities (green lines), and we could not identify distinct structures or “features” in the STRFs at longer than several hundred milliseconds. This suggests a role of A1 neurons in more than detecting stimulus features (Nelken et al., 2003; Nelken, 2004), but it remains to be addressed how neurons in A1 exploit stimulus history and its context on such a long time-scale.

We also examined the performance of nonlinear generic models—i.e., artificial neural networks and support vector regression—and asked if they could reach the upper bound estimated from our experimental data. Because of a limited data length, we examined models with a rather short window length (up to 500 msec; for details, see Section 3.4.2), but even a generic nonlinear model did not reach the upper bound (black lines in Figure 3.11; see also Table 3.2 for models with 200 msec window). One reason for this failure would be an inappropriate choice of initial transformation from time domain to time-frequency domain (and preprocessing). Inspired by the auditory system, here we examined two types of initial transforms—log-frequency spectrograms and cochleagrams in dB scale (Slaney, 1993)—but neither worked well as inputs to artificial neural networks (see also Gill et al., 2006). Another possible mode of the failure would be a rather coarse discretization resolution over time and frequencies (4 msec bin and 3 bins/octaves, respectively), and thus relevant features for neurons might have been lost. Finer discretization however requires a larger amount of data because of an increase in the number of free parameters. Then it would be needed to either employ state-of-the-art methods or devise some judicious nonlinear transformations to extract “relevant features,” and interpret them from biological viewpoints (see also Section 3.6.4).
Figure 3.10: **Poor prediction performance of classical linear-nonlinear cascade models.**

(A) An example of the STRF (200 msec window with $\Delta t = 4$ msec; 18 frequency bins ranging from 0.4 to 22 kHz with $\Delta x = 3$ bins/octave) for an auditory cortical neuron (for details on the estimation, see Section 3.4.1). We can clearly see that the STRF has a broadband excitatory region following a relatively narrow-band inhibitory region. **(B)** Response estimation example. Estimated responses on a validation data set (STRF model as shown in (A), red; Linear-Nonlinear (LN) model, green) capture fairly well the occurrence of depolarization and/or hyperpolarization, but the size is far smaller than the original response (blue). The mean lower and upper model performances in this example were 0.07 and 0.17 for the STRF model, respectively, and 0.10 and 0.22 for the LN model, respectively. **(C)** Relation between model performance and data length. We randomly chose a subset of data to limit the data length (15, 30, 60, and 120 sec; 20 cells from Machens et al., 2004), and assessed the performance of the STRF-based model (200 msec window with the bin size of 4 msec, frequency range of 0.4–22 kHz with discretization resolution 3 bins/octave; STRF, red band for Eq.(3.24); LN, green lines for Eq.(3.27)) computed as in Eq.(3.29). Crosses and open circles respectively show the average model performance on the validation and training data sets, corresponding to the lower and upper estimates of the performance. Model performance does not depend much on the data length, suggesting that the regularization (ridge regression; see Appendix Section A.2.1) worked well, but the optimal regularization strength (the optimal value of $\lambda$ in Eq.(A.53) on page 154) depends on data length in general. The symbols “[” in red and “]” in green at ~40 sec respectively represent the mean performance of STRF and LN models on the data recorded in this study (lower and upper “hems” representing lower and upper estimates, respectively; 23 cells examined with natural sound ensembles), showing that the two data sets from this work and the previous one are comparable. This result justifies the comparison between the upper bound performance (estimated from our data; blue curve) and the model performance (data from Machens et al., 2004; red, green, and black curves) in Figure 3.11.
Figure 3.11: **Context-dependence of response predictability.** Using responses to natural sound probe stimuli in different natural sound contexts (305 probe stimuli in 39 cells), we computed the ratio between the context-independent fraction of the response power and the stimulus-related response power in A1 neurons (*blue*; Eq.(3.23) in Section 3.3.3). The latter is given by the mean response over trials in each context—or the best response estimate under additive noise assumption—and the former is given by the mean over all contexts—or the best estimate of the responses to a probe stimulus without any knowledge on the conditioning stimulus. Therefore, the ratio represents the upper-bound of the response prediction performance for a given window length, which asymptotically approached the true upper limit (*black dotted* line; model performance = 1) by extending the window length—or available stimulus history—on the time-scale of seconds (*cyan*; $\alpha_1 = -0.49$ and $\tau_1 = 1.04$ sec for exponential curve fit as in Eq.(3.22) with $\alpha = 1$). In contrast, the performance (Eq.(3.29); 20 cells from Machens et al., 2004) of linear encoding models (STRF, *red*; Eq.(3.24)) was low for any window length up to $\sim$4 sec even with static nonlinearities (LN, *green*; Eq.(3.27)), and no model could reach the estimated upper bound (ANN, *black*; see also Section 3.4.2) as far as we tested. *Crosses* and *open circles* respectively show the average model performance on the validation and training data sets, corresponding to the lower and upper estimates of the performance. Here we varied the bin sizes for changing the window length of the STRF models while fixing the model complexity (for details, see Section 3.4.1). See also Figure 3.10.
Table 3.2: **Performance of generic nonlinear encoding models.** We examined classical STRF-based models as well as generic nonlinear models having the window length of 200 msec. The inputs to the models were log-frequency spectrograms (except for one case; *ANN with cochleagram*) with rather coarse time and frequency discretizations (for details, see Section 3.4), and the outputs were subthreshold dynamics of auditory cortical neurons from previous works (20 cells; Machens et al., 2004). The model performance (Eq.(3.29)) on the training and validation data sets represents the upper and lower estimates, respectively (see also Figure 3.11). No models reached the upper-bound estimate (∼0.6; Figure 3.11). Acronyms: STRF, spectro-temporal receptive fields; ANN, artificial neural networks; SVR, support vector regression.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear (STRF)</td>
<td>0.17±0.13</td>
<td>0.11±0.09</td>
</tr>
<tr>
<td>Linear-Nonlinear</td>
<td>0.23±0.15</td>
<td>0.14±0.14</td>
</tr>
<tr>
<td>ANN (spectrogram)</td>
<td>0.29±0.14</td>
<td>0.25±0.12</td>
</tr>
<tr>
<td>ANN (cochleagram)</td>
<td>0.27±0.09</td>
<td>0.15±0.10</td>
</tr>
<tr>
<td>ν-SVR</td>
<td>0.32±0.16</td>
<td>0.07±0.12</td>
</tr>
</tbody>
</table>
3.6 Discussion

We have examined how neural responses in the rat primary auditory cortex (area A1) and their predictability depend on stimulus history and its context. Whole-cell recordings performed \textit{in vivo} in the anesthetized rat showed that the context-dependence in A1 sometimes lasted as long as four seconds in some neurons (Figures 3.4 and 3.5); and that the dependence was more susceptible to the changes in lower-order conditioning sound properties—such as overall intensities and frequencies—than to the changes in higher-order properties—such as AM and FM rates and depths—even though the latter had as distinct effects as did the former (Figures 3.8 and 3.9). We also identified that the faithful prediction of response dynamics in A1 requires nonlinear integrations of stimulus history and its context over seconds (Figure 3.11). These results suggest complex mechanisms for stimulus-dependent bottom-up modulation of neural responses in A1, with roles not only in stimulus feature detections but also in stream segregation and other auditory tasks where longer temporal processings would presumably be involved.

3.6.1 Context-Dependence

Our main goal in this study is to characterize the “memory” length of auditory cortical neurons for developing a simple—yet plausible—encoding model well-tailored to the observed properties of response dynamics. Here we used two measures for analyzing context-dependence because we wanted to examine not just whether the effects of past events were statistically significant—i.e., our \textit{ability} to induce and detect context-dependence reliably—but also fractional response power depending on stimulus history—i.e., the \textit{contribution} of the context-dependence effects to response dynamics in A1—as prior knowledge for model construction (for details, see Section 3.3). Both measures identified a slow time-scale in the population data (on the order of seconds; Figure 3.5), while the former also had faster components (on the order of several hundred milliseconds). This is because the neural responses fluctuated a lot
soon after the transition from conditioning to probe stimuli (Figure 3.5B, top panel), presumably as a manifestation of context-dependence induced by our stimuli. The absence of the fast components in the fractional power measure then suggests that the slow time-scale is most relevant from a modeling perspective (see also Figure 3.11), and in fact the time-scale agrees well with that of “echoic memory” (typically between 0.5 and 2 seconds; Neisser, 1967; Glucksberg and Cowen, 1970; Darwin et al., 1972; Triesman and Rostron, 1972; Rostron, 1974; Kubovy and Howard, 1976). Interestingly, in contrast to the monotonic decays in the population data (Figure 3.5), a subset of cells showed nonmonotonic effects that cannot be explained simply by exponential processes (e.g., Figure 3.4). This suggests that the mode of context-dependence could differ from cell to cell. However, we did not have enough data to characterize the relevant time-scales of individual cells, and we conclude that the memory of A1 neurons typically lasts for at least one or several seconds.

Our experimental design differs from previous work in the following three respects. First, most studies on context-dependence thus far used rather simple stimuli, such as pure or AM/FM tones and clicks (e.g., Abeles and Goldstein, 1972; Hocherman and Gilat, 1981; Phillips, 1985; Calford and Semple, 1995; Brosch and Schreiner, 1997, 2000; Ulanovsky et al., 2003, 2004; Bartlett and Wang, 2005; Wehr and Zador, 2005), whereas we used rather complex natural and synthetic stimuli because A1 neurons are extremely sensitive to small changes in stimuli (Bar-Yosef et al., 2002). Second, we sequentially presented such complex stimuli in a randomly interleaved manner instead of just changing inter-stimulus intervals as in conventional two-tone stimulus paradigms. Finally, we examined the context-dependence effects at the subthreshold level to achieve a high temporal resolution despite low firing rates in A1 (Wehr and Zador, 2005; Hromádka, 2007). This novel approach allowed us to detect long-lasting effects that have not been observed in previous studies, and the relation between the time-scales and the stimulus properties we identified (Figures 3.8 and 3.9) could be used as a hint to build a model characterizing the relationship between input acoustic signals and output neural responses. From this modeling viewpoint, it is also advantageous to work on the sub-
threshold levels; such encoding models target the inputs to the cells and thus can be simpler to some extent because we do not have to consider nonlinear spike generating mechanisms of the target cell itself (see also Section 3.6.4 below).

### 3.6.2 Response Predictability

Context-dependence is highly related to response predictability. In fact, the performance of neural encoding models was bounded by the window length, and asymptotically approached the upper limit on the time-scale of seconds ($\tau = 1.04$ sec; Figure 3.11). This long time-scale then explains in part why classical linear encoding models have failed (Sahani and Linden, 2003; Machens et al., 2004). Such models typically have a window length of several hundred milliseconds, but not more than around two-thirds of the stimulus-related predictable response power can be captured on such a short time-scale.

In this study we extended the window length for the linear models (while fixing the model complexity) and incorporated static nonlinearities as well, but the performance did not improve much (Figure 3.11) and we failed to identify stimulus features in STRFs longer than several hundred milliseconds. This casts a doubt on considering neurons in A1 simply as stimulus feature detectors (or, spectro-temporal “edge” detectors; Fishbach et al., 2001, 2003). Many acoustic signal processing problems such as stream segregation in fact require dynamic/nonlinear integration of stimulus history and its context over seconds (Bregman, 1990). The context-dependence reported in this study might then be a neural substrate of such processings, and thus not only system identification but also functional viewpoints might be needed to understand and model the response dynamics in A1.

The findings on the relation between the context-dependence and sound properties (Figures 3.8 and 3.9) imply that such integrations would be somewhat less complex for longer time-scales—i.e., incorporating mainly the lower-order sound properties. However, it remains to be addressed how neurons exploit the sound components on longer time-scales, and it is a
challenge for future works how we could build a plausible “forward” encoding model for A1 neurons.

3.6.3 Possible Mechanisms

Even though our main goal here is not to discover the underlying mechanisms of context-dependence but to characterize the time-course of auditory sensory memory in single-neuron response dynamics, it is always worth speculating the underlying mechanisms.

In contrast to much previous work (e.g., Abeles and Goldstein, 1972; Hocherman and Gilat, 1981; Phillips, 1985; Brosch and Schreiner, 1997, 2000; Bar-Yosef et al., 2002; Ulanovsky et al., 2003, 2004; Bartlett and Wang, 2005), here we examined context-dependence at the subthreshold level and observed significant effects on the time-scale of seconds (Figure 3.5). Although we cannot exclude the possibility that spike generation mechanisms were involved (Carandini and Ferster, 1997; Sanchez-Vives et al., 2000a,b; Gollisch and Herz, 2004), we would then expect that the observed dependence was caused by some cortical network effects (e.g., sensory memory) and by the mechanisms operating at the inputs to the neuron, such as synaptic depression or facilitation (Abbott et al., 1997; Tsodyks and Markram, 1997; Wehr and Zador, 2003; Zhang et al., 2003; Tan et al., 2004; Wehr and Zador, 2005). Such mechanisms could affect different parts of the dendrites, resulting in stimulus-specific effects presumably responsible for the observed nonmonotonicity in some cells (e.g., Figure 3.4; note also stimulus-specific adaptations reported by Ulanovsky et al., 2003, 2004).

We could also speculate that the observed effects would be mainly governed by stimulus-dependent “bottom-up” modulations—because we recorded from anesthetized rats, the semantics of presented sounds for rat behaviors or the state of the animal such as stress levels or attentions would have no effect in this study—and mainly cortical in origin because subcortical auditory neurons have typically much shorter time-scales than cortical neurons (Creutzfeldt et al., 1980; Miller et al., 2002). Then one interesting property of A1 potentially in close relation to the context-dependence is that neurons in A1 typically have short latencies
and precise timing of spiking activities for shorter time-scales (∼10 msec; Phillips and Hall, 1990; DeWeese et al., 2003; Elhilali et al., 2004), while for rather longer time-scales (∼1 sec) auditory cortical neurons are often considered to be “labile” because they tend to adapt quickly to the statistics of stimulus ensembles, or to consecutive presentations of the same stimulus (Condon and Weinberger, 1991; Malone et al., 2002; Ulanovsky et al., 2003, 2004). Small differences in stimulus history would then be accumulated together with response history, and later result in a large difference in the neural behavior where long-lasting effects would emerge in a chain reaction manner. However, such avalanche-like mechanisms could not explain well the nonmonotonic context-dependence effects we sometimes observed, and it remains to be addressed how and where the memory of stimulus history and its context is stored and retrieved.

### 3.6.4 Plausible Encoding Model

Successful curve fitting generally requires appropriate choices of the following three frameworks that allow us to formulate appropriate optimization problems (in the framework of maximum a posteriori inference; see Appendix Section A.2.2): the model class, the loss/cost functions (noise distribution), and the regularizer (the prior). Here we discuss possible ways to develop better—and more plausible—encoding models with respect to these three points.

Two distinct classes of encoding models were examined in this chapter—i.e., classical STRF-based models and generic nonlinear models—but neither of them worked well as shown in Figure 3.11 and Table 3.2. One way to improve the models would be to incorporate observed properties of response dynamics, such as the slow adaptations (τ∼1 sec) identified here. For example, we could employ some adaptive models (with a “forgetting factor” in Eq.(A.61) in Appendix Section A.2.3 (page 157) as: $\xi \sim \exp[-\Delta t/\tau]$; see also Stanley, 2002). Another possible way would be to use recurrent models—such as Kalman filters (Kalman, 1960a,b; Roweis and Ghahramani, 1999), hidden Markov models (Baum and Petrie, 1966; Baum et al., 1970; Rabiner, 1989), and integrate-and-fire-like models (e.g., Paninski et al., 2004, 2005)—to exploit all the stimulus (and response) history of a target cell. Considering that even generic
nonlinear models failed to reach the upper bound estimate of the model performance, however, we would rather attribute the failure to inappropriate choices of the other two frameworks (see below): the loss/cost functions and the regularizer.

In this chapter, all models were trained under the assumption of additive i.i.d. Gaussian noise \((\varepsilon_{ij}(t) \sim \mathcal{N}[0, \sigma_{\text{noise}}^2])\), from Eq.(3.12) on page 97). Our data by and large support this assumption (Figure 3.2 on page 95), but we could use a different noise model—e.g., Laplacian or Cauchy distributions—because the noise distribution of our data was slightly more kurtotic than the Gaussian distribution (\textit{black} vs. \textit{gray} lines in Figure 3.2C, respectively). It might also be interesting to have a different perspective to fit a model to data, e.g., using the maximum informative dimension technique (Sharpee et al., 2004, 2006).

Regularization is critical in solving many problems in science because they are often ill-posed and/or data-limited. Here we imposed a power constraint (or equivalently a Gaussian prior; see Appendix Section A.2.1) on the STRF parameters, but it is not clear if this assumption would be the best for solving the neural encoding problem. One interesting observation is that the distribution of membrane potential is not symmetric but skewed in the positive (depolarization) direction (Figure 3.2A). It would then be interesting to devise or search for methods to exploit this “prior knowledge” for fitting a given class of models.

An alternative approach would be to transform the output responses into appropriate formats, in such a way that the transformed outputs have somewhat simpler—or “more linear”—relations to the input spectrograms. For example, because subthreshold dynamics consist of linear combinations of postsynaptic potentials that have stereotypical shapes in A1 (or “bumps;” Deweese and Zador, 2006)—be they stimulus-evoked or spontaneous—it would be reasonable to identify a template of the bumps and predict the position and scales of the template instead of predicting the membrane potential dynamics themselves. Such templates might be identified by feature extraction methods, and appropriate decompositions—or “bumpgrams”—might be obtained by, e.g., matching pursuit methods (Mallat and Zhang, 1993).