Linearity of cortical receptive fields measured with natural sounds

Christian K. Machens, Michael S. Wehr, and Anthony M. Zador

Cold Spring Harbor Laboratory
One Bungtown Rd
Cold Spring Harbor, NY 11724

Abbreviated title: Linearity of cortical receptive fields

Statistics: 
# text pages: 34  # figures: 7  # tables: -
# words (abstract): about 250 (limit=250)
# words (introduction): about 420 (limit=500)
# words (discussion): about 770 (limit=1500)

Correspondence: Anthony M. Zador
Cold Spring Harbor Laboratory
One Bungtown Rd
Cold Spring Harbor, NY 11724
Tel: 516-367-6950
Fax: 516-367-8866: zador@cshl.edu

Keywords: natural stimuli, auditory cortex, whole cell, patch clamp, reverse correlation, neural code

Acknowledgments: This work has been supported by grants to AMZ from the Packard Foundation, the Sloan Foundation, the NIH and the Mathers Foundation.

Date: September 6, 2003
Abstract

How do cortical neurons represent the acoustic environment? This question is often addressed by probing with simple stimuli such as clicks or tone pips. Such stimuli have the advantage of yielding easily interpreted answers, but have the disadvantage that they may fail to uncover complex or higher-order neuronal response properties. Here we adopt an alternative approach, probing neuronal responses with complex acoustic stimuli, including animal vocalizations. We have used in vivo whole cell methods in the rat auditory cortex to record subthreshold membrane potential fluctuations elicited by these stimuli. Most neurons responded robustly and reliably to the complex stimuli in our ensemble. Using regularization techniques, we estimate the linear component—the spectro-temporal receptive field (STRF)—of the transformation from the sound (as represented by its time-varying spectrogram) to the neuron’s membrane potential. We find that the STRF has a rich dynamical structure, including excitatory regions positioned in general accord with the prediction of the simple tuning curve. However, while the STRF successfully predicts the responses to some of the natural stimuli, it surprisingly fails completely to predict the responses to others; on average, only 11% of the response power could be predicted by the STRF. Most of the neuron’s response, therefore, cannot be predicted by the linear component, even though the response is deterministically related to the stimulus. We analyze the systematic errors of the STRF model qualitatively and show that the underlying nonlinear response properties are mostly dynamic, i.e., are due to the time-varying properties of the neural encoder.
1 Introduction

While it is widely agreed that the primary visual cortex decomposes images into components such as oriented edges (Hubel and Wiesel, 1962), the corresponding decomposition of acoustic stimuli in the primary auditory cortex remains uncertain. The spectro-temporal receptive field (STRF) has recently attracted increased interest as a candidate framework for characterizing the function of auditory cortical neurons (Kowalski et al., 1996a; Kowalski et al., 1996b; deCharms et al., 1998). Part of the appeal of the STRF rests in its conceptual simplicity; like its successful visual counterpart, the spatio-temporal receptive field, it offers a straightforward linear description of a neuron’s behavior.

The cortical STRF has been estimated using a variety of stimulus ensembles, including dynamic ripples (Kowalski et al., 1996a; Klein et al., 2000; Miller et al., 2002), random chord stimuli (deCharms et al., 1998; Rutkowski et al., 2002; Sahani and Linden, 2003), and random tone pips (Blake and Merzenich, 2002). However, the ultimate test of any model of sensory function rests in the ability of the model to predict responses to natural stimuli. It remains at present an open question how well the STRF can explain the behavior of the auditory cortex under natural conditions, where an organism encounters highly complex, dynamically changing stimuli. While natural stimuli have long been used to probe cortical responses (Wollberg and Newman, 1972; Sovijärvi, 1975; Creutzfeldt et al., 1980; Nelken et al., 1999), and have been widely used in other preparations to compute STRFs (Eggermont et al., 1983; Yeshurun et al., 1989; Theunissen et al., 2001), they have only rarely been used to test the STRF on cortical neurons (Rotman et al., 2001; Machens et al., 2003).

Here we estimate the STRF defined by subthreshold membrane potentials using in vivo whole cell recording. The output of cortical neurons is sparse—a point process...
consisting mostly of zeros (silence) punctuated by the occasional one (a spike). Sub-threshold responses, by contrast, are analog—the membrane potential can in principle assume any value within some range given by the various ionic driving forces—and so provide a much richer source of information about the neuron’s response. In addition, the subthreshold membrane potential permits insight into the computations performed by the total synaptic input to a neuron. It should be noted that whole cell recording also has a different sampling bias from conventional extracellular recording: instead of recording from active neurons with large action potentials, (i.e., those that are most easily isolated on the electrode), whole cell recording selects for neurons solely on the basis of the experimenter’s ability to form a gigahm seal.

Using these novel methods, we have analyzed the response properties of single neurons in the primary auditory cortex (region A1) of rats. In particular, we have focused on two questions: What kind of STRFs do we obtain using subthreshold responses recorded in whole-cell mode? How well do these STRFs predict the responses of cortical neurons to natural sounds?

2 Methods

2.1 Surgery

Sprague-Dawley rats (p17–20) were anesthetized with ketamine (60 mg/kg) and medetomidine (0.48 mg/kg) 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 that left the ears free and clear. Local anesthetic was applied to the scalp, and a small craniotomy and durotomy were performed above the left auditory cortex. A
cisternal drain was performed prior to the craniotomy. Prior to the introduction of electrodes, the cortex was covered with physiological buffer (in mM: NaCl, 127; Na₂CO₃, 25; NaH₂PO₄, 1.25; KCl, 2.5; MgCl₂, 1) mixed with 1.5% agar. 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.

2.2 Whole-cell recordings

We used standard blind whole-cell patch clamp recording techniques, modified from brain slice recordings (Stevens and Zador, 1998). Membrane potential was sampled at 4 kHz in current clamp ($I = 0$) mode using an Axopatch 200b amplifier (Axon Instruments, Union City, CA) with no online series resistance compensation. Electrodes were pulled from filamented, thin-walled borosilicate glass (1.5 mm O.D., 1.17 mm I.D.; World Precision Instruments, Sarasota, FL) on a vertical 2-stage puller. Internal solution contained, in mM, K-gluconate 140, HEPES 10, MgCl₂ 2, CaCl₂ 0.05, MgATP 4, NaGTP 0.4, Na₂Phosphocreatine 10, BAPTA 10, QX-314 (an intracellular sodium channel blocker for blocking action potentials) 5, Alexa-594 (a fluorescent dye) 0.1; pH 7.25; diluted to 290 mOsm. Mean series resistance was 82.9 ± 16.5 MΩ, and mean resting membrane potential was $-70.0 ± 8.8$ mV ($N = 22$). Resistance to bath was 3–5 MΩ before seal formation.

Recordings were made from primary auditory cortex (A1) as determined by the tonotopic gradient and “V-shaped” frequency-amplitude tuning properties of cells and local field potentials. We recorded from the superficial layers (subpial depth range: 203–526 µm, as determined from micromanipulator travel). One cell was recovered histologically; it was verified to be a layer 2/3 pyramidal cell. Altogether, we recorded
from \( N = 22 \) cells. Some neurons \((N = 3)\) responded so rarely that they did not allow the computation of the linear response component (see Figure 1).

### 2.3 Stimulus presentation

Pure tone stimuli (frequencies: 1 to 40 kHz in 1/3 octave increments; attenuations 10 to 70 dB in 20 dB increments, with a peak amplitude of 66 dB SPL) were sampled at 97.656 kHz and had a duration of either 25 msec with 5 msec 10–90% cosine-squared ramp, or 70 msec with 20 msec ramp, and were delivered in a pseudorandom sequence at a rate of 1–2 Hz. The stimulus level was calibrated using a Bruel and Kjaer 4939 1/4” microphone (Bruel and Kjaer, Norcross, GA).

All natural sounds were taken from commercially available audio CDs, sampled at 44.1 kHz. Sound sections of animal vocalizations were selected from “The Diversity of Animal Sounds” and “Sounds of Neotropical Rainforest Mammals” (Cornell Laboratory of Ornithology, Ithaca, NY). A variety of sound sections of environmental noises were taken from the CD series “Great Smoky Mountains National Park” (Cornell Laboratory of Ornithology, Ithaca, NY) and “Spectacular Sound Effects” (Madacy Records, Montreal, Canada). The beginning sequence of *Purple Haze* by Jimi Hendrix was taken from audio CD. While the majority of the sound sections lasted for 7.5-15 sec, some were considerably shorter (down to 1 sec for the sound of a closing door, for instance) while some lasted longer if they were deemed to have sufficient complexity (31 sec for Jimi Hendrix, for instance).

Altogether, our ensemble of natural stimuli consisted of more than \( n = 120 \) different sounds. The stimuli covered all frequencies from 0–22 kHz and ranged from narrow-band stimuli (such as cricket calls) to broad-band stimuli (such as a gurgling creek). Figure 2A (black line) shows the average power spectrum of the natural stimuli.
tested on the cells in this study. Note that only a subset of the natural stimuli was tested on any particular cell so that the power spectra usually differed from cell to cell. This subset was chosen so that significant power fell into the range of frequencies covered by a neuron’s tuning curve. The red lines in Figure 2A indicate the spread (here measured as the standard deviation) of these power spectra. The distribution of modulation amplitudes (here defined as the square root of the power measured in short time bins, $\Delta t = 1 \text{ ms}$) of the natural stimuli is displayed in Figure 2B. About 15% of the time, the natural stimuli featured relative quiet (more than 70 dB below the peak amplitude). The power spectrum of the amplitude modulations is shown in Figure 2C.

For stimulus presentation, the natural sounds were resampled at 97.656 kHz. All stimuli were delivered using a System 3 Stimulus Presentation Workstation (Tucker-Davis Technologies, Alachua, FL) at 97.656 kHz to a calibrated electrostatic speaker and presented free field in a double-walled sound booth. Stimuli were presented with a peak amplitude of 66 dB SPL.

In a first set of experiments ($N = 10$), a fixed subset of natural sounds was used and repeated up to 20 times. These experiments allowed us to assess response reliability. In a second set of experiments ($N = 12$), as many natural sounds as possible were presented, each once or twice only.

Some of the natural stimuli are referred to in the text and figures. The abbreviations used are as follows: JC = Jaguar Mating Call (*Panthera onca*), BW = Bowhead Whale (*Balaena mysticetus*), KF = Knudsen’s Frog (*Leptodactylus knudseni*), BM = Bearded Manakin (*Manacus manacus*).  

---

6
2.4 Data analysis

All data analysis was performed in MATLAB (MathWorks, Natick, MA). Responses to conventional pure tone stimuli were assessed by constructing frequency-intensity profiles, in which the evoked membrane potential for each frequency and intensity was averaged across trials (Figure 1A,B). Frequency tuning curves at a given intensity were constructed using the peaks of these mean evoked responses (Figure 7E,F). Best frequency (BF) was defined as the frequency which evoked the maximal mean membrane potential at a given intensity, whereas characteristic frequency (CF) was defined as the frequency at which a response could be evoked at the lowest possible intensity.

The responses to natural stimuli were analyzed by means of the STRF. As a first step of the data analysis, all natural stimuli were transformed into the time-frequency domain using the spectrogram method (Cohen, 1995; Klein et al., 2000). This transformation is used as a rough approximation of the cochlear transform and considerably simplifies the subsequent analysis of auditory computation. Use of the spectrogram also facilitates comparison with other studies in the field. The spectrogram method depends on a particular choice of a time window. For a given window function \( h(\tau) \), the energy density spectrum of the sound pressure wave \( s(t) \) is given by (Cohen, 1995)

\[
P(t, f) = \left| \frac{1}{2\pi} \int d\tau e^{-2\pi i f \tau} s(\tau) h(\tau - t) \right|^2 .
\]

Here we used the Hamming window function (Press et al., 1992).

The numerical analysis requires a discretization of both time and frequency. To account for properties of the cochlea, we used a logarithmic discretization of the frequency axis. Within a reasonable and computationally feasible range (time window and discretization \( \Delta t = 5–25 \) ms; frequency discretization \( \Delta f = 1–5 \) frequencies/octave), several choices were used independently of each other, essentially yielding the same
results. Given equally spaced time steps \( t_i \) with \( i = 1 \ldots M \) and logarithmically spaced frequency steps \( f_l \) with \( l = 1 \ldots L \), we computed the discretized spectrogram \( S(t_i, f_l) \) as

\[
S(t_i, f_l) = 20 \log \left[ \int_{t_i}^{t_i + \Delta t} dt \int_{f_l}^{f_l + \Delta f} df P(t, f) \right]. \tag{2}
\]

To estimate the response, the stimulus spectrogram \( S(t_i, f_l) \) was filtered linearly with the spectro-temporal receptive field \( H(-t_k, f_l) \) of the neuron,

\[
\hat{r}(t_i) = r_0 + \sum_{k=1}^{K} \sum_{l=1}^{L} H(-t_k, f_l) S(t_i - t_k, f_l), \tag{3}
\]

where \( r_0 \) is a constant offset. (We use a negative time index in the STRF, \(-t_k\), for formal equivalence with the conventions of the reverse correlation approach.) Assuming a finite memory of the system, the STRF has a finite temporal extent as indicated by the indices \( k = 1 \ldots K \). We distinguish the estimated response \( \hat{r}(t_i) \) from the measured response \( r(t_i) \) by a hat. Note that the response is usually taken to be the average firing rate (Eggermont, 1993; Klein et al., 2000; Theunissen et al., 2000), whereas here the response is given by the subthreshold voltage trace.

To estimate the STRF, we used linear regression techniques which generalize the more widely used reverse correlation methods to arbitrary stimulus ensembles (Klein et al., 2000). To illustrate this approach, it is helpful to simplify the notation. We will write \( \hat{r}_t = \hat{r}(t_i) \) and re-order indices such that \( a_j = H(-t_k, f_l) \) and \( s_{ji} = S(t_i - t_k, f_l) \) with \( j = (l-1)K + k \). Furthermore, without loss of generality, we center both response and stimulus to have zero mean, \( \langle \frac{1}{M} \sum_i r_i \rangle = 0 \) and \( \langle \frac{1}{M} \sum_i s_{ji} \rangle = 0 \) for all \( j \), where angular brackets denote averaging over trials. It follows that \( r_0 = 0 \) so that the model (Eq. 3) simplifies to

\[
\hat{r}_t = \sum_{j=1}^{N} a_j s_{ji}. \tag{4}
\]
where $N = KL$. By definition, the STRF is now given by the parameters $a_j$ which can be fitted by minimizing the mean-square error between the estimated response $\hat{r}_i$ and the measured response $r_i$:

$$\text{Err} = \frac{1}{M} \sum_{i=1}^{M} \left[ r_i - \sum_{j=1}^{N} a_j s_{ji} \right]^2$$

This is the problem solved by multi-dimensional linear regression. In terms of the stimulus-response cross-covariance, $A_k = \frac{1}{M} \sum_{i=1}^{M} s_{ki} r_i$ and the stimulus-stimulus auto-covariance $B_{jk} = \frac{1}{M} \sum_{i=1}^{M} s_{ji} s_{ki}$, the solution is given by

$$a_j = \sum_{k=1}^{N} B^{-1}_{jk} A_k$$

The negative power denotes the matrix inverse. In the neurophysiological jargon, $A_k$ is usually referred to as the reverse correlation function and $B_{jk}$ as the auto-correlation of the stimulus. “White” stimuli are often defined as stimuli whose auto-covariance matrix corresponds to the identity matrix. In these cases, the reverse correlation function equals the STRF.

For natural stimuli, use of the auto-covariance matrix is crucial in order to divide out the stimulus correlations, Eq. 6 (Theunissen et al., 2001). However, additional complications may arise from undersampling, if the number of stimulus-response pairs is too small to obtain an adequate estimate of all the coefficients of the STRF. Mathematically, this problem is reflected in an auto-covariance matrix that has many eigenvalues close to zero. The inversion of this matrix therefore results in a very noisy estimate of the STRF, corresponding to strong overfitting of the poorly sampled dimensions and poor predictive power of the model.

To address this issue, we used a regularization approach, which places constraints
on the parameter values (Hastie et al., 2001). We used two types of constraints. The first penalizes strong deviations of the parameters from zero; this is the same constraint used in ridge regression (Hastie et al., 2001). The second penalizes strong deviations between neighboring parameters, and therefore enforces smoothness of the STRF: Consequently, the parameters $a_j$ are obtained by minimizing the following error function,

$$\text{Err} = \frac{1}{M} \sum_{i=1}^{M} [r_i - \sum_{j=1}^{N} a_j s_{ji}]^2 + \lambda \sum_{j=1}^{N} a_j^2 + \mu \sum_{j=1}^{N} \sum_{k \in \mathcal{N}_j} (a_j - a_k)^2$$

(7)

where $\mathcal{N}_j$ denotes the set of indices describing the neighbors of $j$. In the toy example shown in Figure 4A, the set is given by $\mathcal{N}_7 = \{2,6,8,12\}$. The parameters $\mu$ and $\lambda$ determine the strength of the constraints. For notational simplification, we write

$$\text{Err} = \frac{1}{M} \sum_{i=1}^{M} [r_i - \sum_{j=1}^{N} a_j s_{ji}]^2 + \sum_{j=1}^{N} \sum_{k=1}^{N} C_{jk} a_j a_k$$

(8)

where the constraints are now absorbed in the matrix elements

$$C_{jk} = (\lambda + 2 |\mathcal{N}_j| \mu) \delta_{jk} - 2\mu \sum_{l \in \mathcal{N}_j} \delta_{lk}$$

(9)

Here $\delta_{lk}$ denotes the Kronecker delta with $\delta_{lk} = 1$ if $l = k$ and $\delta_{lk} = 0$ otherwise; $|\mathcal{N}_j|$ denotes the number of elements in the set $\mathcal{N}_j$. The minimization of Equation 8 with respect to $a_j$ now results in

$$a_j = \sum_{k=1}^{N} (B + C)^{-1}_{jk} A_k$$

(10)

In the case $C_{jk} = \lambda \delta_{jk}$, this solution reduces to ridge regression (Hastie et al., 2001) and in the case $C_{jk} = 0$ to “naive” regression, Eq. 6.

So far, the method leaves the two constraint parameters $\mu$ and $\lambda$ undetermined.
When the data are split into a training and a prediction set, the STRF can be estimated on the training set for fixed values of the constraint parameters. This STRF can in turn be used to estimate the responses in the prediction set. Repeating this procedure for different constraint parameters, we find the parameters that minimize the mean square error between the estimated and actual responses of the prediction set, see Figure 4B-E. Given data for \( n \) natural stimuli, we used \( n - 1 \) stimuli for the training set and the remaining data for the prediction set. To avoid overfitting of the constraint parameters on a specific prediction set, this procedure was repeated on permutations of prediction and training sets; in the end, the average constraint parameters were selected for final evaluation of STRFs and prediction errors.

### 2.5 Analysis of prediction errors

The utility of the STRF is ultimately determined by its ability to estimate correctly the response. A natural measure for the quality of these estimates is the mean square error between estimated and measured response, \( \sigma^2_r = \langle \frac{1}{M} \sum_i (r_i - \hat{r}_i)^2 \rangle \) where the angular brackets denote trial averaging. However, even for a perfectly linear system, this error will not be zero because the response is contaminated by a certain amount of noise (Sahani and Linden, 2003). Hence, the best we can do is reach this level of noise. Assuming a simple additive model of response and noise, the residual noise component can be estimated as

\[
\sigma^2_\eta = \frac{N}{N-1} \left[ \frac{1}{M} \sum_i r_i^2 - \frac{1}{M} \sum_i \langle r_i \rangle^2 \right]
\]

where \( N \) is the number of trials. Given the response power \( \sigma^2_r = \langle \frac{1}{M} \sum_i r_i^2 \rangle \), a natural measure of the relative success of the STRF model is given by (Sahani and Linden,
which generally varies between 0% (when the mean square error $\sigma^2_e$ equals the response variance) and 100% (when the mean square error reaches the residual noise $\sigma^2_\eta$). On the training set, the relative success might also exceed 100% in the case of overfitting. For the same reason, the relative prediction success can fall below 0%.

In cases where only one trial was available ($N = 6$ cells), the noise component could not be computed from the data. In order to compare estimates of the training and prediction success in these cells, we set the noise component in the one-trial experiments to a conservative 50% of the response power; this fraction corresponds to the relative noise level measured in the least reliable cells.

To uncover temporal structure of the error function, we resolved the error in the frequency domain, using the coherence function defined as (Brockwell and Davis, 1991)

$$
\gamma(f) = \frac{|S_{rr}(f)|}{S_{rr}(f)S_{\hat{r}\hat{r}}(f)}
$$

(13)

where $S_{rr}(f)$ is the Fourier transform of the cross-correlation between $r_i$ and $\hat{r}_i$—also termed the cross-spectrum—and $S_{rr}(f), S_{\hat{r}\hat{r}}(f)$ are the power spectra of measured and estimated responses, respectively. The coherence takes values between zero (no correlation between measured and estimated response at a certain frequency) and one (perfect correlation).

## 3 Results

In this study, we sought to characterize cortical neurons on the basis of their responses to natural sounds. In particular, we tested the ability of a linear model to account for
the stimulus-response relation. Our analysis consisted of the following steps. First, we quantified responsiveness and response reliability. Next, we computed the linear component of the stimulus-response relation (the STRF). Finally, we quantified the ability of this linear component to approximate the actual responses of auditory cortical neurons, and characterized the successes and failures of this linear predictor.

3.1 Tuning curves and responsiveness

We recorded intracellularly from single neurons in primary auditory cortex of rats using the whole-cell in vivo patch recording technique. We prevented action potentials pharmacologically using the intracellular sodium channel blocker QX-314 (see Methods), so that recordings consisted only of fluctuations in the subthreshold membrane potential—the total synaptic input to the cell, prior to thresholding by the spike generating mechanism. We emphasize that the absence of spikes implies that any non-linearities in the stimulus-response relationship cannot be attributed to the effect of spike threshold in the neuron under study.

While most neurons featured strong subthreshold membrane potential fluctuations, a few neurons were essentially unresponsive to natural stimuli, except for transient onset responses to any sound. For these cells, unresponsiveness to natural stimuli co-existed with robust and reliable responses to conventional pure tone stimuli presented at 1-2 Hz. Figure 1 compares the responses of two cells to pure tones and to an animal vocalization. While both cells showed robust responses to pure tones (Figure 1A,B), with similar frequency and intensity tuning, one cell (Figure 1E) responded strongly to the natural sound while the other cell (Figure 1F) did not. The natural stimulus shown in this example contained power at the preferred frequencies of both cells (Figure 1G), so neither the tuning of the cells nor the spectral structure of the stimulus can easily
explain the striking difference in responsiveness of these two cells. Altogether, we found a continuum of responsiveness across cells, as measured by the square root of the average power in the responses (Figure 1I). Note that this measure takes into account both stimulus-locked and stimulus-independent activity.

Although the existence of such unresponsive neurons is intriguing, establishing the causes underlying this strong adaptation is beyond the scope of this study. Because the absence of stimulus-locked fluctuations in membrane potential in unresponsive neurons precluded the estimation of reliable STRFs, these neurons ($N = 3$) were excluded from our analysis and are not included in the results below.

### 3.2 Natural stimuli

The natural sounds used in our study were mostly animal communication calls and environmental sounds that lasted for 7.5–15 seconds. The sounds were chosen for their spectral and temporal complexity and diversity, rather than on the basis of neuroethological considerations. The overall ensemble consisted of more than $n = 120$ different sounds, of which only a subset was tested on any particular cell. This subset was chosen to match approximately the frequency tuning of the cells. Figure 2A shows the average power spectrum of these stimulus subsets, demonstrating that ample power fell into the frequency range covered by most of the cells (compare Figure 1H).

The distribution of modulation amplitudes (Figure 2B) and the power-law behavior of the power spectrum of these amplitudes (Figure 2C) are in accordance with earlier observations on the statistics of natural sounds (Attias and Schreiner, 1997). The left peak in Figure 2B corresponds to silence in the sounds (more than 70 dB below peak stimulation); the stimuli were relatively silent for about 15% of the time.

The spectrograms of three example sections of natural sounds used in this study are
3.3 Response reliability

Responsive neurons typically showed a combination of both spontaneous and stimulus-locked voltage fluctuations in response to natural stimuli (Figure 1E, 2E). Both spontaneous and stimulus-locked responses are presumably due to the synchronous arrival of many postsynaptic potentials (PSPs). If spikes had not been blocked pharmacologically, the larger PSPs would likely have triggered spikes. With at most 2–3 large PSPs per second, the activity of the recorded neurons is temporally sparse.

Neurons sometimes showed striking trial-to-trial reliability. This is particularly evident in the central panel of Figure 2E where the responses to repeated presentations of the same stimulus are nearly identical. Reliability was stimulus dependent: The same neuron was less reliable for a different stimulus (see Figure 2E, right panel).

To quantify the amount of stimulus-locked activity, we compared a single response trace with the average over the remaining trials. A sample comparison (Figure 3A, same data as in Figure 2D,E, left panel) shows that the deviations of a single trial from the average mostly involved the fine structure of the voltage fluctuations. To quantify this observation, we computed the coherence function between the single and the average trace. The coherence measures the frequency-resolved correlation of two time series (see Methods), and ranges from zero (absence of stimulus-locked activity) to one (when all traces feature the same stimulus-locked excursion in membrane potential). The average coherence functions corresponding to the three examples of Figure 2D,E are shown in Figure 3B. These functions demonstrate the typical range of stimulus-independent background activity observed in the experiments. All cells feature reliable activity for lower frequencies (<40 Hz). However, when presented with the “right”
stimulus, the coherence increased dramatically: The light gray curve (BM) shows the coherence corresponding to the central panel in Figure 2D,E.

Response reliability also differed from cell to cell. Figure 3C displays the average magnitude of the stimulus-independent activity. To compute this quantity, the variance of the response about its mean was averaged over time (see Methods). In all cases, the average magnitude of the noise (1–5 mV) is small compared to the magnitude of the PSPs (10–30 mV), emphasizing the overall reliability of the responses.

### 3.4 Spectro-temporal receptive fields

In the next step, we characterized the linear component of the stimulus-response relation. This task is considerably simplified when the stimulus is represented by a spectrogram (Cohen, 1995; Klein et al., 2000) as in Figure 1C,D and Figure 2D. The spectrogram provides a rough approximation of the processing performed by the cochlea, as the sound pressure wave is transformed into a cortical response. Using the spectrogram representation of the stimulus, we then analyzed the linear component of the response, i.e., the spectro-temporal receptive field (STRF).

The STRF has often been estimated using the reverse-correlation method (Eggermont, 1993; deCharms et al., 1998) based on well-defined random stimuli. Natural stimuli, however, feature correlations in both the temporal and spectral domain. Linear regression generalizes this approach to arbitrary stimulus ensembles by dividing the reverse correlation solution—technically the cross-covariance between the stimuli and the response—by the auto-covariance of the stimulus, Eq. 6 (Theunissen et al., 2001).

Additional complications may arise from undersampling, i.e., if the number of stimulus-response pairs is too small to obtain an adequate estimate of all the coefficients of the STRF. To avoid overfitting along the undersampled directions, we devel-
oped a procedure that incorporated power and smoothness constraints on the STRF parameters (see the Methods section for details).

In general, either the power or the smoothing constraint was sufficient, with little predictive power gained by combining them. However, the trade-offs between these constraints permitted us to assess the robustness of the STRF estimates; STRFs computed from the same set of data with different constraints are shown in Figure 4C and D. We found that major features of the STRFs, such as the approximate positions of the excitatory and the inhibitory peaks, were generally independent of the precise constraints used. Minor characteristics, such as the relative widths of these peaks, were however more sensitive to the precise details of the regularization. In what follows, we report the STRFs obtained using the smoothing constraints, as these could more easily be interpreted at a glance, but we emphasize that because not all characteristics were robust, care must be taken to avoid over-interpreting the details of the STRF structure.

STRFs typically featured an arrangement of both inhibitory and excitatory fields as shown by a few examples in Figure 5A–D. The excitatory (red) and inhibitory (dark blue) fields indicate times and frequencies at which stimulus energy leads to, respectively, an increase or decrease in the neuron’s response. Because an inhibitory field usually preceded an excitatory field, the STRF often predicted strong responses to stimulus onsets within a specific frequency range, as was in fact observed. Excitatory regions usually extended about 1–3 octaves and 50–100 ms.

While qualitatively similar STRFs have been reported for spiking neurons in similar preparations (deCharms et al., 1998; Klein et al., 2000), our STRFs appear to be both temporally and spectrally more extended. This is consistent with previous observations that subthreshold tuning curves to pure tones are broader than the corresponding suprathreshold tuning curves (DeWeese and Zador, 2000).
Within the reliability of our estimates, most cells had STRFs (such as the ones shown in Figure 5A–D) that were approximately separable (Depireux et al., 2001); that is, these STRFs can be re-written in the form \( H(t, f) = \Theta(t)\Phi(f) \) where the functions \( \Theta(t) \) and \( \Phi(f) \) now completely describe the time- and frequency components.

The STRFs derived from natural stimuli are generally consistent with the neuron’s frequency sensitivity as measured with short pure tones (see Methods). The frequency sensitivity of the STRF can be obtained by plotting the maximum STRF values for every frequency; a comparison of the derived and measured curves is shown in Figure 5E and F. While the overall frequency sensitivity is captured by the STRF, the curves nevertheless differ in their details. These differences may arise in part from a fundamental limitation of all linear models, which require that tuning curves obtained at different intensities may differ only in magnitude and not in form; no linear model can account for an amplitude-dependent shift in best frequency. Comparison of the pure tone tuning curves measured at 46 vs. 66 dB SPL (Figure 5E and F, green and blue curves) shows that their forms differ; note the shifts in the best frequency and the appearance of a second peak in the 66 dB responses. Such intensity dependence is effectively averaged out in the STRF model. Nevertheless, across the population the best frequencies derived from the STRFs were in rough agreement with those measured by pure tones (Figure 5G).

### 3.5 Stimulus dependence of training success

Although the linear component of the input-output function of most neurons showed well-defined structure, the analysis presented thus far does not indicate how strong this linear component is, i.e., how much of the stimulus-locked response we can capture using the (purely linear) STRF alone. To assess the potential limitations of the lin-
ear STRF, we tested it on the set of data that was used in its training. To compare different natural stimuli, the data used in the estimation of the STRFs were restricted to 15-second-long stimulus sections (most of which were individual stimuli) and their respective responses. The relative training success was quantified as the percentage of the stimulus-locked response variance captured by the STRF (see Methods). This procedure yields an upper bound for the quality of any linear model.

As indicated in Figure 6, we found considerable variance in our data with some recordings that led to a good training success (up to 75%) and some recordings that fared considerably worse (down to 2%). The latter is particularly surprising as it shows that the linear model can fail completely. Note that these failures cannot be trivially attributed to differences in the amount of stimulus-locked activity, since the training success in Figure 6 is quantified with respect to the amount of stimulus-locked activity only (see also Methods). Interestingly, the variance in the data can partially be explained by a systematic stimulus-dependence of the training success: Some stimuli (such as JC, see red dots in Figure 6) always lead to a high training success while others (such as BW, green dots in Figure 6) are constantly worse.

Even though we used regularization, the training success is artificially boosted due to some overfitting. In particular, certain stimuli might lend themselves to stronger overfitting and lead to a systematically higher training success than other stimuli. To clarify these issues, we also investigated how well the STRF can predict the responses to stimuli not included in its estimation. Given an individual stimulus and its respective responses, the STRF was estimated on the remaining data and its prediction on the individual stimulus was evaluated. This procedure yields the relative prediction success which is compared against the relative training success in Figure 7. The relative prediction success provides a lower bound on the performance of the STRF model. As
demonstrated by Figure 7, the differences between individual stimuli are retained for the relative prediction success.

Altogether, the relative success of the linear STRF model is bounded between 11\% (the average relative prediction success) and ???\% (the average relative training success). The “true” relative success of the linear model has to lie within these bounds.

### 3.6 Qualitative characterization of the non-linearities

The widespread failure of the linear model to predict responses for many but not all complex stimuli indicates a high but stimulus-dependent degree of nonlinearity. By comparing the predicted and actual responses, we can attempt a first qualitative characterization of the non-linearities involved.

Three sample predictions are shown in Figure 7A-C for the neurons in Figure 2D,E and 3A,B. While the predicted trace (red line) in Figure 7A accounts for the approximate times at which PSPs occur, it does not capture their precise shape. This observation can be quantified by spectrally resolving the prediction success. For that purpose, we again use the coherence function as a measure of the correlation at each frequency, in this case between actual and predicted response (Figure 7D, solid line). Clearly, this particular STRF does not predict any response fluctuations faster than \( \approx 5 \) Hz. As a comparison, recall that the response is reliable up to at least 20 Hz (dashed line in Figure 7D).

Figure 7B shows a natural stimulus that elicited a highly reliable response which the STRF failed to predict. The example uses the same data as in Figure 2D,E, central panel. While the STRF predicts the timing of the PSPs, it underestimates their amplitudes (see arrows in Figure 7B). These situations are quite common in the data and might be explained by adaptive mechanisms that modify the gain of the STRF on longer time scales. (???)
Many of the STRFs predicted strong responses to the onset of stimuli (ON-responses) within a certain frequency range (Figure 5A-D). Accordingly, these STRFs also predict strong negative responses at stimulus termination. As can be seen in Figure 7C (right arrow), the cell’s actual response is sometimes just the opposite: a strong excitatory response at the termination of the stimulus, termed an OFF-response. Because the OFF-response has the same sign as the ON-response, it represents a form of rectifying nonlinearity.

Figure 7C also demonstrates that the linear model can completely fail to predict the occurrence of PSPs (left arrow); such failures lead to a correspondingly weak coherence (Figure 7F). These cases might be attributed to dynamic interactions between ON- and OFF-responses, depending on their spectro-temporal relations (Tai and Zador, 2001) or other (unknown) non-linearities.

3.7 Ruling out trivial non-linearities

The extent to which the linear model fails is quite surprising. This raises the question whether there might be some “trivial” effects at work.

For instance, static non-linearities (due to saturation or rectification as mentioned above) could possibly explain part of the failures. By plotting the predicted response against the actual response in a calibration plot, these systematic errors can be visualized, cf. Fig. 7G–I. The rectification due to OFF-responses shows up as dots in the upper left quadrant, see Figure 7H and I. However, the comparatively small number of points in this quadrant illustrates that OFF-responses are not a major factor in the prediction error. Accordingly, any static non-linearity fit to the calibration plot and incorporated into the model does not significantly enhance the prediction success.

Previous tests of linearity in cortical neurons have assumed the neurons to be in an
adapted state. Natural sounds are non-stationary stimuli, meaning that their statistics (such as mean intensity) fluctuate over time. The permanent changes in mean intensity might therefore rule out that the neuron settles in an adapted state. Correspondingly, this might deteriorate the estimates of the STRF and be responsible for systematic differences between stimuli. To see whether this is the case, we computed the variations in the mean intensity of different stimuli and plotted it against the prediction success. Figure ??A shows an example of such a scatter plot for variations of the mean intensity in time windows $T = 6$ seconds. Independent of the chosen time window, there is almost no correlation between the STRF’s ability to estimate the response (training success) and the fluctuations in mean intensity of the stimuli. Hence, adaptation to mean intensity is only a negligible factor in the failure of the STRF to predict the responses to natural stimuli.

The failure of the STRF model therefore has to be attributed to other forms of non-linearities. These might include adaptation to other parameters of the stimulus (e.g. auditory contrast) and the above mentioned two-tone interactions.

4 Discussion

We have used whole cell patch clamp methods in vivo to record subthreshold membrane potential fluctuations elicited by natural sounds. In the majority of cells, subthreshold responses were sufficiently rich and robust to permit a reliable estimation of the linear predictor of the neuron’s response, the STRF. The present article represents the first analysis of subthreshold responses elicited by natural stimuli in vivo.

Major response properties such as frequency tuning were similar whether assessed by pure sine tones or complex sounds. The STRFs estimated from complex sounds, however, provided a much more complete view of the neuron’s dynamics, so that it was
possible to compare the predicted and experimentally measured responses to complex stimuli.

Prediction success depended strongly on the particular sounds used in the experiment (Figure 6). On average only about 11% of the response power could be predicted by the STRF, indicating that the neuron’s response was highly nonlinear. These nonlinearities are mostly dynamic, and included both adaptation and second order interactions. The presence of these strong non-linearities should also caution the reader against over-interpreting the STRF, since non-linearities in responses will create artificial structure in the linear STRF. ¹

Our observations are in accord with recent work on neurons in the auditory forebrain of zebrafinches (Theunissen et al., 2000), where neurons show a high degree of feature selectivity in response to natural stimuli. On the other hand, previous work in the auditory cortex of ferrets has suggested that the responses of cortical neurons can be well predicted by the linear STRF (Kowalski et al., 1996b; Klein et al., 2000; Schnupp et al., 2001). However, our results have shown that the success of the STRF depends strongly on the type of stimulus used (see Figure 6). Ripple stimuli (and combinations thereof) could therefore fall into the class of stimuli for which responses can be well predicted by a linear model. It will be interesting to investigate whether the STRF model also provides good predictions in this system when more complex stimuli are used.

A recent study by Sahani and Linden (2003) confirms the presence of non-linearities in the rat auditory cortex. Using random chord stimuli, they showed that about 40% of the stimulus-locked response power could be explained by the linear STRF model. In

¹WHAT ARE THE POSSIBILITIES FOR A LINEAR MODEL TO STILL ACCOUNT FOR DATA? MUST BE VERY RAGGED STRF (OURS ONLY SMOOTHED VERSION). HIGHLY UNLIKELY BECAUSE OF TUNING CURVE
contrast, the linear model could only account for 11% of the stimulus-locked response power in our study. The use of stationary random stimuli versus non-stationary natural stimuli might again explain the differences in these findings.

The concept of the STRF, derived and evolved from the second-order Volterra kernel has long been used as a tool in auditory research (Eggermont, 1993). Unfortunately, this history has led to various definitions of the STRF. In our definition, the STRF constitutes a linear transform between the spectrogram of the stimulus and the response; this definition is thus similar in spirit to the work of Kowalski et al. (1996b) and deCharms et al. (1998). Recent work has also fitted a second-order Volterra series to the responses of neurons in the auditory cortex of anesthetized cats (Rotman et al., 2001). Although natural stimuli were used in the estimation and although a second-order kernel acting on the sound pressure wave is formally equivalent to the linear STRF model acting on the spectrogram, there are important differences from our work: Rotman et al. used very short snippets of natural stimuli (125 msec length) to estimate correspondingly short kernels (6 msec length of prediction); for comparison recall that we used 7.5–15-second-long stimuli and STRFs which were 250 msec long. These large differences render direct comparison difficult.

The estimation of STRFs from natural stimuli presents a statistical challenge as these stimuli fill the high-dimensional stimulus space in an irregular fashion. Conventionally, researchers have sought to reduce the dimensionality of the problem. One possibility is to expand the STRF or second-order Volterra kernel in a few number of basis functions (Yeshurun et al., 1989). Another possibility is to restrict the stimulus to basis functions derived from the principal components of the natural stimuli (Theunissen et al., 2001). In the statistical literature, the latter is sometimes referred to as principal component regression (Hastie et al., 2001).
Here we used a more general approach based on regularization techniques that constrain the model parameters without any prior dimensionality reduction. A similar approach has recently been taken to compute STRFs of simple cells in V1 from natural stimuli (Smyth et al., 2003). Recent developments also seek to solve the estimation problems using evidence optimization (Sahani, 2003) or by fitting the best linear stimulus subspace (as opposed to the best linear model) to the neurons (Sharpee et al., 2003; Paninski, 2003).

In the end, however, we believe that the most urgent problem concerns the quantitative characterization of the observed non-linearities. Explaining these nonlinearities represents an exciting challenge for future research.

References


Figure 1: Responsive and unresponsive cells. We used in-vivo whole-cell methods to record subthreshold responses of single neurons in auditory cortex A1. Action potentials were blocked with the intracellular sodium channel blocker QX-314 (see Methods). A,B, Responses of two cells to conventional pure tone stimuli. Evoked membrane potentials are shown for an array of frequencies and intensities (the loudest tones are on the top row). Both cells exhibited robust responses to pure tones, with typical “V-shaped tuning,” and had similar characteristic frequencies (CFs) of 3.2 kHz (A) and 4 kHz (B). C,D, Spectrogram of a 5-second segment of the call of a Knudsen’s Frog (stimulus KF). E,F, Responses of these two cells to this sound were strikingly different. In E, this stimulus evoked robust and reliable responses, whereas in F, after a transient onset response, the cell was completely unresponsive. The cell in F was similarly unresponsive to all six natural stimuli tested (not shown). G, This stimulus contained power at the CFs of both cells (arrows show CFs, colors match traces in A,B,E,F). In fact, stimulus power was greater at the CF of the unresponsive cell. H, Most cells in our sample had CFs of 1-5 kHz. Arrows show CFs of the two cells in A,B,E,F. I, Responsiveness to natural stimuli varied across cells. Here responsiveness is quantified by the standard deviation of the membrane potential evoked by natural stimuli (note that non-stimulus-evoked activity also contributes to this measure). Arrows show the different responsiveness of the two cells in E,F.

Figure 2: Natural stimuli and responses. Most of the stimuli employed in this study were animal communication signals and environmental noises. A, Power spectrum of natural sounds. Usually, the sets of natural sounds tested on different cells varied slightly. The figure shows the mean (black line) of the power spectra of these different ensembles as well as their standard deviation (red lines). B, Distribution of modulation amplitudes, same format as in A. The large peak on the left corresponds to moments of
relative silence in the stimuli. C, Power spectrum of the modulation amplitudes, same format as in A. D, Spectrograms of three short stimulus sections. The spectrograms illustrate some of the diversity and complexity of the natural signals employed in this study. E, Subthreshold membrane potential responses. The traces show highly reliable stimulus-locked activity to ten repetitions of the corresponding stimuli as well as some spontaneous events. The level of spontaneous activity was stimulus-dependent; note the very low level in the central panel.

**Figure 3:** Reliability of responses. A, Mean response compared to a single trial for a natural stimulus (same data as in Figure 2D,E, left panel). The overall correspondence between the two traces shows that the amount of spontaneous activity is relatively small. B, Average coherence functions between the mean response and a single trial for the data shown in Figure 2D,E. The curves demonstrate that the average level of background activity depends on the stimulus. C, Noise level for different cells. The noise level was quantified as the average deviation of the single response trials from the mean response (see Methods). While the noise level differed between cells and stimuli, it was always small compared to the size of the PSPs which typically ranged between 10-30 mV.

**Figure 4:** STRF estimation. To estimate a relation between stimulus and response, we computed the spectro-temporal receptive fields (STRFs) of the recorded neurons. To circumvent estimation problems deriving from the usage of natural stimuli, we subjected the STRF estimation to a smoothing and a power constraint. A, Smoothing constraint. The smoothing constraint enforces that the values of neighboring bins do not deviate too strongly. The neighbors of bin 7, for example, are shown in gray. B, Naïve estimate of the STRF via linear regression. An estimate without any constraints achieves a mean square error $\varepsilon = 5.6 \text{ mV}^2$ between actual and predicted re-
response on the data used for the STRF estimation (training) and an error $\epsilon = 10.69 \text{ mV}^2$ on new data (prediction). The large difference indicates strong overfitting which is also visible in the noisy structure of the STRF. C, Optimal estimate of the STRF subject to a power constraint. Here, the power constraint was chosen to minimize the prediction error. Indeed, while the training error increases ($\epsilon = 6.95 \text{ mV}^2$), the prediction error is now considerably lower ($\epsilon = 9.97 \text{ mV}^2$). D, Optimal estimate of the STRF subject to smoothing constraint. Both training error ($\epsilon = 7.08 \text{ mV}^2$) and prediction error ($\epsilon = 10.01 \text{ mV}^2$) are similar to those in C. E, Prediction error for different combinations of smoothing and power constraints. For this cell, combining the two types of constraints does not significantly enhance the prediction success. Hence, all STRFs inside the trough (blue contours) are equally valid estimates; showing the “extremes” in B and C allows an assessment of the robustness of the estimates.

Figure 5: STRFs and tuning curves. A-D, STRFs for four different neurons. The STRFs feature both negative (“inhibitory”) and positive (“excitatory”) contributions to the response, here displayed by dark blue and yellow-red colors, respectively. All STRFs show a sequence of inhibitory and excitatory fields; this characteristic predicts positive responses to sound onsets. E,F, Tuning curves. The STRFs predict a specific frequency tuning, here shown as solid red lines for the STRFs in C (panel E) and D (panel F). Overall, this prediction is in accord with the frequency sensitivity measured with pure tones. For comparison, two tuning curves recorded at 66 dB SPL (blue) and 46 dB SPL (green) are displayed. G, Comparison of best frequencies as measured by the tuning curve and the STRF. The conventional tuning curve exhibit a range of best frequencies at different intensities, here displayed as black bars. The characteristic frequency (best frequency at the lowest intensity) of the cells is shown as a red cross. Overall, the best frequencies predicted by the STRF (abscissa) are in good agreement
with those measured with pure tones (ordinate). Some cells with incomplete tuning curves were excluded.

**Figure 6:** Summary of prediction success. The main figure shows a scatter plot of training versus prediction success. Each point represents the prediction success of the model on novel data (Y-axis) and the corresponding success on training data (X-axis). To permit the individual points on the graph to be resolved, only a subset of the stimuli ($n = 10$; chosen randomly for each neuron) are shown. While the prediction success provides a lower bound on the capability of the model to estimate the response, the training success yields an upper bound. Surprisingly, some of the stimuli are consistently better than others across neurons, compare JC (red dots) and BW (green dots). Hence, the STRF is able to capture a significant part of the response to some stimuli, yet it fails to predict the response to others. The distribution of training and prediction success is displayed as histogram on the top and on the right, respectively. Averaged over all stimuli and cells, the training success was 39% and the prediction success was 11% (arrows). Hence, the responses of cortical neurons to natural stimuli are dominated by nonlinearities.

**Figure 7:** Prediction success and failures. A–C, Spectrogram, measured and predicted responses for the same data as shown in Figure 2. In A, the prediction (red) captures the gross features of the mean response (black), but not the fine details. In B, the STRF rightly predicts the occurrence of most PSPs but markedly fails to predict their overall size (arrows). In C, the STRF not only underestimates the size of PSPs but at times completely fails to predict their overall occurrence (left arrow), hinting at more complicated nonlinearities. Note in particular the last PSP (right arrow) where the cell responds to the termination of a stimulus (OFF-response)—here the prediction has the wrong sign, a case of rectifying nonlinearity. D–F, Coherence between measured and
predicted responses (solid lines), corresponding to the data shown in A–C, respectively. The coherence functions underpin the observation that the STRF succeeds at best in capturing slower temporal components. For comparison, the dashed lines replot the coherence between a single trial and the mean (compare Figure 3B) which provide an upper bound. G–I, Calibration plot, same data as in A–C, respectively. Plotting the predicted versus the actual response reveals any static, systematic errors inherent to the linear model. The black lines show the baselines of the actual responses. While the plot in G suggests an overall linear relationship between actual and measured responses, the plots in H and I demonstrate the presence of nonlinearities. The largest part of the prediction error is due to the failure of the STRF to predict the correct size of the PSPs.

**Figure 8**: Effects of mean intensity variations on the training success. A, Scatter plot of the standard deviation in mean intensity (in $T = 6$ second time windows) in different stimuli versus the respective training success. There is only a very weak correlation. B, Correlation coefficients for different time windows $T$. Fluctuations in the mean intensity of different stimuli are not an important factor in the failure of the STRF model.