Supporting Information

Kayser et al. 10.1073/pnas.1012656107

SI Materials and Methods

Correction of the Limited Sampling Bias in Direct Estimates of Information. Direct estimates of mutual information can suffer from a bias due to the limited amount of data available to calculate the conditional probabilities $P(r|s)$ (compare Eq. **1**). This sampling bias in the information estimate was corrected by using a multistep procedure, which combines the established shuffling technique (1) and the quadratic extrapolation procedure (2). For details of this procedure, see refs. 3 and 4. The performance of this procedure with the number of trials was tested on simulated data in several previous reports (3–6). A test of the performance on simulated data with first-order statistical properties matching those of the real auditory cortical neurons analyzed here was performed as follows (we refer to ref. 3 for full details): We computed the time varying peristimulus time histogram (PSTH) in 4-ms bins collected in response to the natural sound stimulus. Then, for a given neuron a simulated set of responses was generated by using an in-homogeneous Poisson process with the same PSTH as in the real experiment. We then studied the variation of the information estimates as a function of the number of simulated trials included in the estimate (see figure S2 in ref. 3 for an example). The critical parameters for the performance of bias correction procedures are the number of trials per stimulus N (which was in the range 39–70 in the real data, with mean 55) and the cardinality R of the response space (i.e., the number of possible different responses that can be observed), which was $R = 2^6 = 64$ in this study. The larger R (and the smaller N), the more difficult is to correct for the bias (1). The simulations showed that our algorithm is highly data robust, with very mild degradation of performance when decreasing the number of trials. In particular, the information estimate converged to within 4 and 1% of the "asymptotic" (large trial number) value when using 32 and 64 trials per stimulus, respectively (average over all simulated neurons), suggesting that our information theoretic calculations provide a good estimate of the actual values.

Evaluation of the Relative Importance of PSTH Modulations and of Correlations Between Spike Times in Information Transmission. In the main text, we computed how mutual information $I(S;R)$ (Eq. **1**) depends on the precision at which neural responses are sampled. This calculation is useful to prove that precise spike times are important for encoding, but it is not sufficient to determine what aspects of the spike train are most crucial for information transmission. The simplest possibility is that the only information-bearing feature of the spike train is the temporal modulation of the PSTH. The PSTH is proportional to the time-dependent firing rate and expresses the first-order statistical properties of the spike train at a given temporal resolution. A second possibility is that neuronal firing cannot be described completely by its first-order statistics, and information is encoded by high order statistical properties of the spike train, such as correlations between spike pairs that cannot be explained by firing rate modulations. In this second case, an observer of the spike train would not be able to decode all information available in the spike train unless it takes into account such higher-order correlations between spike times.

We characterized the relative importance of PSTH modulations and correlations between spike times by estimating how much stimulus information can be extracted even by a decoder that considers only the time-dependent firing rate and ignores temporal correlations. By following refs. 7–9, a lower bound to this information can be computed by using the following equation:

$$I_{\text{PSTH}} = \sum_{r,s} P(r|s) \log_2 \frac{P_{\text{PSTH}}(r|s)}{P_{\text{PSTH}}(r)}, \quad \text{[S1]}$$

where $P_{\text{PSTH}}(r|s)$ is the probability of observing response r to stimulus s if the neuron fired according to a time-dependent Poisson process with the same PSTH as the one of the neuron under analysis, and $P_{\text{PSTH}}(r) = \sum_r P_{\text{PSTH}}(r|s)P(s)$.

To determine whether correlations are needed to extract most information from the spike train, it is useful to compare the value of I_{PSTH} to the value of $I(S;R)$, which quantifies the overall amount of information that can be obtained from the neuronal response taking into account all its characteristics. By information theoretic inequalities it can be proved that $I_{\text{PSTH}} \leq I(S;R)$. If the ratio $I_{\text{PSTH}}/I(S;R)$ is close to 1, then most information carried by the spike train can be extracted even by a downstream observer of neural activity that pays attention only to the first-order statistics and neglects correlations. The quantity I_{PSTH} was corrected for the limited sampling bias using the same procedures used for the overall information $I(S;R)$ and reported above.

Stimulus Decoding. An alternative approach to computing the mutual information between stimulus and neural response are methods based on explicit models for stimulus decoding, such as linear discriminant analysis (10, 11). By making particular assumptions, such as the linear separability of the responses to different stimuli, these methods reduce the statistical complexity of estimating the information conveyed by neural responses about the stimulus and can yield reliable results also for problems where only small amounts of experimental data are available.

However, a decoding approach may miss critical ways in which a response eliminates uncertainty about the stimulus, such as by making incorrect or oversimplified assumptions about the stimulus–response relationship. For example, a linear decoder (as e.g., described below) would fail to capture the contribution of nonlinear stimulus–response relationships. Another reason why decoding approaches may not capture all information encoded by a neural response is that responses may encode information by other means than just indicating the most likely stimulus (11, 12). For example, the knowledge that a particular stimulus is totally unlikely can provide considerable information that would not be captured by a decoder extracting the most likely stimulus, even if this decoder makes all of the proper assumptions about the stimulus–response relationship (11). In contrast to the decoder, information theory has the advantages of quantifying all of the knowledge that can be gained from a response and of taking possible nonlinear stimulus–response relationships into account.

Here, we used a linear decoder to probe the stimulus–response relation at different response precisions. Details of the procedure are described as follows.

General principle. Stimulus decoding analysis is a direct approach to test how well a predetermined set of stimuli can be discriminated given the set of observed responses. Here we used a linear discriminant decoder in conjunction with a leave-one-out cross-validation procedure. Practically, the decoding was done as follows. We randomly selected eight 1-s chunks from each of the long sounds (Fig. S2) and used these as eight stimuli for decoding. For each trial and stimulus, we then repeated the following: (*i*) The average responses to all other seven stimuli were computed across all repeats of the respective stimuli. (*ii*) For the current stimulus the mean response was computed by averaging across all trials, excluding the current "test" trial. (*iii*) The Euclidean distance (over time points) was computed be-

tween the response on the test trial and all eight average responses. The test trial was decoded as that stimulus yielding the minimal distance to the test response. This procedure was repeated for each trial of each of the eight stimuli, and the total percentage of correctly decoded trials and the confusion matrix were determined. It should be noted that this decoding operation is based only on the likelihood of the neural response given the stimulus and, therefore, implicitly assumes that all stimuli have the same prior probability of being presented, as it actually happens in our experiment by design.

Response definition and temporal precision. Decoding was applied to the same eight stimulus chunks used for the information analysis and individual neuron's responses were sampled at 1-ms resolution (resulting in a response vector of 250 points). To study the impact of temporal response precision on decoding, we used a temporal shuffling procedure. The temporal precision of the response was manipulated by shuffling (independently across trials and stimuli) the responses in groups of n neighboring time bins, where the parameter n indicates the temporal extend of the shuffling. Importantly, this shuffling procedure preserves the original number of time bins and, hence, the statistical dimension of the data, but reduces the effective precision at which the response is "read" by the decoder.

Decoding performance quantification. To quantify the performance of the decoder, we derived the information provided by the "confusion" matrix of the decoder. The values on a given row s and column d of the confusion matrix $Q(d|s)$ represent the fraction of trials on which the presented stimulus s was decoded to be stimulus d. If decoding is perfect, the values in Q will be one on the diagonal and zero otherwise. The confusion matrix is of interest not only because it provides an intuitive picture of the decoding success and errors, but also because it provides a direct link to information theory. The information in the confusion matrix is defined by the following equation:

$$I(S;D) = \sum_{d,s} P(s)Q(d|s)\log_2\frac{Q(d|s)}{Q(d)}. \qquad \textbf{[S2]}$$

Information theoretic inequalities ensure that $I(S;D) \leq I(S;R)$, with $I(S;R)$ being the direct information estimate (13). It is important to note that even for an optimal decoder, the extracted information may be strictly less than the information available in the response. This inequality comes because the decoding operation captures only one aspect of the information carried by the population response, namely the identity of the most likely stimulus. Mutual information [$I(S;R)$], in contrast, quantifies the overall knowledge about the stimulus gained with the single-trial response, including information carried by the absence of a response in a particular time window, or by providing evidence about the relative likelihood of different stimuli (14). As a consequence, methods of (linear) decoding can provide considerable

insight about the stimulus–response relationship but may fall short in capturing all of the knowledge that can be gained by observing a neural response.

Characterization of Neural Precision by Peristimulus Time Histogram (PSTH) Autocorrelation Width. By following Desbordes et al. (15), the temporal precision of a neuron's response can be measured by using the width of the central peak in the temporal autocorrelogram of the trial averaged PSTH. Under assumptions described in ref. 15, this analysis provides a good measure of the width of typical peaks in the neuron's response. We here computed this measure by first calculating the temporal autocorrelation of the trial averaged response and defined the response precision as the half-width at half-height (HWHH) of the central peak of the autocorrelogram.

SI Results

Stimulus Decoding Versus Direct Information Estimates. Several previous analyses on the time scales of stimulus representations were based on decoding techniques. By making specific assumptions, such as the linear separability of the responses to different stimuli, these methods reduce the statistical complexity and can yield reliable results also when only small amounts of experimental data are available. However, even for optimal well constructed decoders, the extracted information may be strictly less than the total information available in the response (11), because the decoding operation captures only one aspect of the response: the identity of the most likely stimulus. Direct estimates of mutual information [$I(S;R)$], in contrast, quantify the overall knowledge about the stimulus gained from a single-trial response including, for example, information carried by the amount of certainty with which the most likely stimulus was predicted, or by providing evidence about which are the less likely stimuli given the neural response. As a consequence, decoding methods might underestimate the contribution of precise spike timing.

Indeed, when using a linear decoder to probe the stimulus–response relation at different effective precisions, we found a reduced information loss compared with the direct information estimates reported above. Fig. S3 directly compares the normalized information values obtained with the direct estimate from the responses to random chords (as in Fig. 1*E*) to the information extracted from the linear decoder (Eq. **S2**). For all temporal precisions, the information loss suggested by the decoder was smaller than that obtained from the direct information estimate, with the difference between the two methods amounting up to 60% at 48 ms effective precision. This result shows that the information carried by precisely timed spikes is better revealed when considering the full dependency between stimulus and response and might be underestimated by decoding approaches.

1. Panzeri S, Senatore R, Montemurro MA, Petersen RS (2007) Correcting for the sampling bias problem in spike train information measures. *J Neurophysiol* 98:1064–1072.
2. Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W (1998) Entropy and information in neural spike trains. *Phys Rev Lett* 80:197–200.
3. Kayser C, Montemurro MA, Logothetis NK, Panzeri S (2009) Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron* 61:597–608.
4. Montemurro MA, Rasch MJ, Murayama Y, Logothetis NK, Panzeri S (2008) Phase-of-firing coding of natural visual stimuli in primary visual cortex. *Curr Biol* 18:375–380.
5. Montemurro MA, Senatore R, Panzeri S (2007) Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Comput* 19:2913–2957.
6. Magri C, Whittingstall K, Singh V, Logothetis NK, Panzeri S (2009) A toolbox for the fast information analysis of multiple-site LFP, EEG and spike train recordings. *BMC Neurosci* 10:81.
7. Nirenberg S, Latham PE (2003) Decoding neuronal spike trains: How important are correlations? *Proc Natl Acad Sci USA* 100:7348–7353.
8. Pola G, Thiele A, Hoffmann KP, Panzeri S (2003) An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network* 14:35–60.
9. Latham PE, Nirenberg S (2005) Synergy, redundancy, and independence in population codes, revisited. *J Neurosci* 25:5195–5206.
10. Nelken I, Chechik G (2007) Information theory in auditory research. *Hear Res* 229:94–105.
11. Quian Quiroga R, Panzeri S (2009) Extracting information from neuronal populations: Information theory and decoding approaches. *Nat Rev Neurosci* 10:173–185.
12. Rieke F, Warland D, de Ruyter van Steveninck RR, Bialek W (1999) *Spikes - Exploring the Neural Code* (MIT Press, Cambridge, MA).
13. Cover TM, Thomas JA (1991) *Elements of Information Theory* (Wiley-Interscience, London).
14. Panzeri S, Treves A, Schultz S, Rolls ET (1999) On decoding the responses of a population of neurons from short time windows. *Neural Comput* 11:1553–1577.
15. Desbordes G, et al. (2008) Timing precision in population coding of natural scenes in the early visual system. *PLoS Biol* 6:e324.
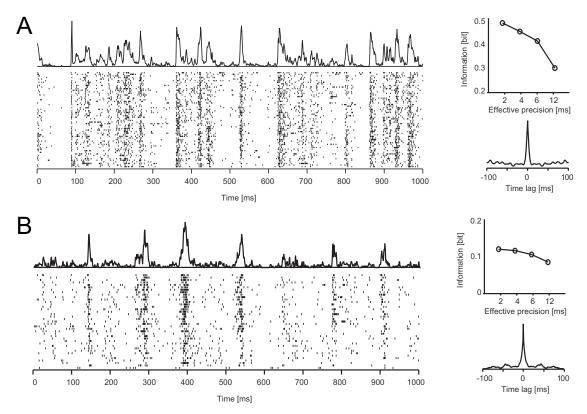
Fig. S1. Responses of example neurons to random chords (*A*) and natural sounds (*B*). (*Left*) Spike times on individual repeats of the same stimulus together with the trial-averaged response (upper trace). (*Right*) The stimulus information obtained at different effective temporal precisions (*Upper*) and the response autocorrelation (*Lower*).
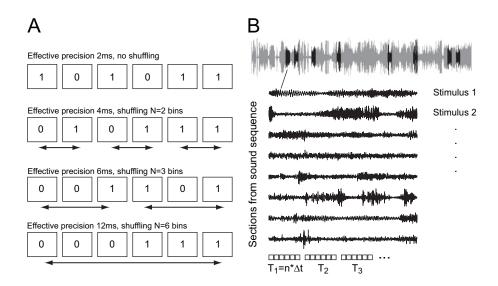
Fig. S2. Details of the temporal shuffling procedure and the stimulus definition. (*A*) Temporal shuffling procedure to reduce the effective response precision. Spikes were shuffled (independently across trials and stimuli) across *N* neighboring time bins. In particular, we shuffled spikes in neighboring bins (doubling the effective resolution, *n* = 2), in triplets (increasing resolution by a factor of three, *n* = 3) or all six spikes in the pattern (increasing resolution by a factor of six, *n* = 6). Boxes correspond to time bins (here, Δt = 2 ms). (*B*) Stimulus definition for information calculation. Stimuli were eight sections (black traces) randomly chosen from a long sound sequence (gray trace). For the quantification of spike patterns, each section was divided into nonoverlapping time windows Ti consisting of *n* = 6 time bins (Δt = 1, 2, 4, or 8 ms). Mutual information was calculated between responses (spike patterns), and these eight stimuli separately for each time window Ti. Information values were averaged over all subsequent time windows to provide the information conveyed by the responses for discriminating the eight different sound sections. To ensure the independence of our results from the eight stimulus sections used to calculate the information values, we repeated the same procedure by using 50 sets of randomly selected sections and averaged the information values.
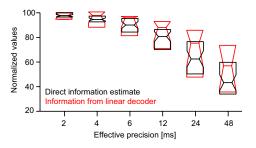

Fig. S3. Comparison of direct information estimates and those derived from stimulus decoding. The figure displays the normalized information values derived from direct estimates of the stimulus information (black, same as in Fig. 1*E*) and from the confusion matrix of a linear decoder (red). It demonstrates that in this situation, a linear decoding method reveals a smaller drop in stimulus information with progressively coarser effective temporal precision than the direct method. For each neuron, information was normalized (100%) to the value at 1-ms precision. Boxplots indicate the median and 25th–75th percentile across neurons.