Journal of Computational Neuroscience 19, 199–221, 2005

Encoding Stimulus Information by Spike Numbers and Mean Response Time in Primary Auditory Cortex

ISRAEL NELKEN*
Department of Neurobiology and the Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem, Israel
israel@md.huji.ac.il

GAL CHECHIK†
Department of Computer Sciences, Hebrew University, Jerusalem, Israel

THOMAS D. MRSIC-FLOGEL‡, ANDREW J. KING AND JAN W. H. SCHNUPP
University Laboratory of Physiology, University of Oxford, Parks Road, Oxford, UK

Received January 9, 2005; Revised March 27, 2005; Accepted April 27, 2005

Action Editor: Matthew Wiener

Abstract. Neurons can transmit information about sensory stimuli via their firing rate, spike latency, or by the occurrence of complex spike patterns. Identifying which aspects of the neural responses actually encode sensory information remains a fundamental question in neuroscience. Here we compared various approaches for estimating the information transmitted by neurons in auditory cortex in two very different experimental paradigms, one measuring spatial tuning and the other responses to complex natural stimuli. We demonstrate that, in both cases, spike counts and mean response times jointly carry essentially all the available information about the stimuli. Thus, in auditory cortex, whereas spike counts carry only partial information about stimulus identity or location, the additional availability of relatively coarse temporal information is sufficient in order to extract essentially all the sensory information available in the spike discharge pattern, at least for the relatively short stimuli ($<\sim$100 ms) commonly used in auditory research.

Keywords: auditory cortex, complex sounds, electrophysiology, mutual information

Introduction

A central question in neurobiology is how neurons encode information in spike trains. Neuronal responses are usually quantified by counting the number of spikes evoked by a set of stimuli during a relatively long time window. This implicitly assumes that important information about the stimuli is encoded exclusively or predominantly by the firing rate. However, stimulus information could also be encoded in other features of the spike train, such as spike latencies, interspike intervals, and so on. In fact, recent work suggests that first spike latencies, a highly reduced form of timing information, may be more informative than spike count or rate

*To whom correspondence should be addressed.

†*Present address*: Computer Science department, Stanford University, Stanford CA 94305, USA.

‡*Present address*: Max Planck Institute for Neurobiology, Am Klopferspitz 18a, 82152 Martinsried, Germany.

in the coding of space by neurons of auditory cortex (Brugge et al., 2001; Furukawa and Middlebrooks, 2002; Jenison and Reale, 2003). The importance of first spike latency in encoding stimulus location or identity has also been emphasized in studies of the visual (Van Rullen et al., 1998; Van Rullen and Thorpe, 2001, 2002) and somatosensory systems (Panzeri et al., 2001; Johansson and Birznieks, 2004).

One way of investigating the importance of these features is to create a 'decoder', which receives a spike train as input and guesses which stimulus evoked the spike train. Various decoders have been proposed, including neural networks (Middlebrooks et al., 1994, 1998) and maximum likelihood estimators (Jenison and Reale, 2003). The performance of these estimators has been quantified in two ways. One way is to estimate their probability of making a correct decision. The other is to generate a 'confusion matrix', and to compute the mutual information (MI, also called transmitted information) between actual stimuli and decoded stimuli, i.e. between the rows and columns of the confusion matrix (Victor and Purpura, 1996; Furukawa and Middlebrooks, 2002).

A constant uncertainty in these studies is that all of the resulting estimates depend on the use of an appropriate decoder. Any given result might, in principle, be improved if a better decoder, capable of reading more information out of the spike trains, could be found. With the resurgence of information-theoretic methods in neuroscience (Rieke et al., 1997), it has been realized that there is an upper bound on the information that even the best decoder can extract, and that this bound is also a measure of MI, but this time of the joint distribution between stimuli and spike trains. The upper limit is set by the information processing inequality (Cover and Thomas, 1991), which states that any function applied to the responses (for example by the decoder as it interprets the spike train) can never increase the information about the stimuli. At best it can preserve the MI, but it will often reduce it.

This absolute limit on the performance of any decoder makes the calculation of the MI between stimuli and spike trains an important tool in studying the neural code. The MI between stimuli and full spike patterns serves as an absolute scale with respect to which the performance of any decoder, or reduced measure, can be judged. By comparing the MI between stimuli and any reduced measure of the response, such as spike counts or latency, with the absolute bound, we can make quantitative statements about how appropriate (or 'informative') this measure is as a putative neural code. Therefore, if the MI of a confusion matrix generated by a decoder is close to the absolute bound, it can be said that this decoder is close to optimal.

The greatest difficulty in estimating the MI carried in spike patterns stems from the fact that the space of possible spike patterns is extremely large, and is therefore only very sparsely sampled by the data. With sparse data, standard MI estimators become both biased and unreliable since they tend to over-fit to the available training data. A number of very different methods have been proposed to overcome this (e.g. Rolls et al., 1997; Panzeri and Schultz, 2001; Treves, 2001; Victor, 2002; Pola et al., 2003).

Here, we directly compare these methods, and apply them to two very different sets of data collected in auditory cortex, one measuring spatial receptive field properties, the other responses to complex natural stimuli (bird song samples). For the purpose of evaluating the methods, we used benchmark simulated data for which the true MI was known. We then applied the method that performed best on the simulated data to the two datasets. We show that, for both datasets, a pair of reduced measures, namely spike count and mean response time, together carry as much information as can be extracted from the full spike pattern by any tested method. Thus, for these datasets, this pair of reduced measures carries all the information about the stimuli present in the neural responses, and can therefore serve as a "complete" neural code.

Methods

Experimental Data

Two sets of data were analyzed here. The first was collected in ferret A1, where the stimuli were broadband noise bursts shaped to mimic sounds from the free-field (virtual acoustic space (VAS) stimuli). These recordings were carried out in Oxford, UK, using protocols approved by the UK Home Office. The details of the stimuli and data collection are described elsewhere (Schnupp et al., 2001; Mrsic-Flogel et al., 2005). In short, ferrets were anesthetized with alphaxalone/alphadolone acetate (SAFFAN; 2 ml·kg^{-1}, i.p., supplementary doses i.v.). Following the completion of surgery, the animal was transferred to a continuous i.v. infusion of pentobarbital sodium anesthetic (SAGATAL; 2–3 mg·kg^{-1}·h^{-1}, i.v.) and gallamine triethiodide (FLAXEDIL; 20 mg·kg^{-1} h^{-1}, i.v.). The animal was ventilated artificially and body temperature, end-tidal CO$_2$, heart rate, electrocardiogram and electroencephalogram were monitored continuously. The left A1 was exposed and single unit activity was recorded using tungsten-in-glass electrodes, TDT System II and BrainWare (Tucker-Davis Technologies, Alachua, FL). We estimated the neural spatial response fields (SRFs) by recording responses to VAS stimuli (20 ms gaussian noise bursts, convolved with the animal's own head related transfer function to provide life-like interaural time, level and spectral cues, and presented over calibrated headphones).

To measure a neuron's SRF, responses to five presentations from each of 224 different virtual source directions were collected (presentations from different directions were randomly interleaved).

The second dataset consisted of responses of cat A1 neurons to bird song stimuli. The details of the stimuli and data collection are described elsewhere (Bar-Yosef et al., 2002). In short, cats were premedicated with xylazine (1 mg, i.m.). Ketamine (30 mg/kg i.m.) was used for induction of anesthesia. The cat was then ventilated with a gas mixture of O$_2$/N$_2$O (30%/70%) with halothane (0.2–1%, as needed). Blood pressure, heart rate, end-tidal CO$_2$, and body temperature were continuously monitored. Recordings were performed in low-frequency A1, as judged by anatomical landmarks and physiological response properties. Recordings were performed using 2–4 tungsten electrodes. The electrodes could be moved independently (EPS, Alpha-Omega, Nazareth, Israel). On-line spike sorters (MSD, Alpha-Omega, Nazareth, Israel) were used to detect spikes. Only well-separated spikes are analyzed here.

The stimulus set was derived from three bird song recordings (these are stimuli 1, 3 and 5 of Bar-Yosef et al., 2002). Each chirp was modified by separating it into the main tonal component (an amplitude- and frequency-modulated tone) and the noise components. The noise was further subdivided into the echoes and the unrelated background. The stimulus set used here comprised the original recording, the main component, main + echo, noise, and the unrelated background of each of the three bird chirps, giving a total of 15 distinct, complex sound stimuli. The stimuli were usually presented at an attenuation level nominally 20 dB above the units' noise threshold. Twenty repetitions of each stimulus were collected, in a pseudo-random order.

General Considerations in Estimating the MI

If we think of the identity of a stimulus S and a neural response measure R as random variables, then the MI between stimulus and response, measured in bits, is a functional of their joint distribution $p(s, r)$, and is defined as

$$I(S; R) = \sum_{s,r} p(s, r) \log_2 \frac{p(s, r)}{p(s)p(r)}, \qquad (1)$$

where $p(s)$, $p(r)$ are the marginal distributions (Cover and Thomas, 1991). The MI is zero if the two variables are independent (that is $p(s, r) = p(s)p(r)$ for every r and s), and is positive otherwise.

The MI measures the tightness of the relationship between stimuli and responses and in this respect is similar to other statistical tests for the sensitivity of a neuron to the stimuli, e.g. 1-way ANOVA on the spike counts with the stimulus as a factor, or a χ^2 test of independence between the histograms of spike counts for the different stimuli. In fact, under the assumption of independence, the MI is asymptotically identical to the χ^2 measure, up to a multiplicative factor, and is therefore χ^2 distributed (Sokal and Rohlf, 1981, pp. 721–724).

MI has some important theoretical advantages over standard statistical tests. In particular, in a χ^2 test, once the null hypothesis (i.e. X and Y are independent, and hence "mutually un-informative") has been rejected, it is difficult to interpret the value of the statistic. In contrast, the MI value has a clear interpretation: it is the logarithm to the base 2 of the number of different stimulus classes that can be distinguished on the basis of the response. For example, if the MI is equal to the entropy (a measure of the initial uncertainty, Cover and Thomas, 1991) of the stimulus set then it is possible to build a perfect (i.e. essentially error-free) decoder of the neural activity. If all stimuli in the dataset are equally likely, then the entropy of the stimulus set is simply equal to the logarithm to the base 2 of the number of stimuli in the set. When the MI is smaller than the stimulus entropy, it can be used to give an estimate of the number of neurons required to build a perfect decoder: this will be close to the entropy divided by the typical MI, with corrections due to informational redundancy or synergy that may occur among the neurons.

Estimating the MI between stimuli and responses accurately is very difficult. Naïve MI estimators may, for different reasons, suffer from both overestimation and underestimation. For the purpose of this paper, the reasons for underestimation are the most important. The MI represents the maximal amount of information that any decoder could possibly extract from the neural activity. This is due to the data processing inequality, which states that for any function f of the responses,

$$I(f(R), S) < I(R, S), \tag{2}$$

where R is the random variable denoting the response and S is the random variable denoting the stimuli. Therefore any transformation of the response R into $f(R)$ can be considered as a data reduction step. The transformation f can represent a decoder such as a neural network, or the derivation of reduced measures such as spike counts or response latency.

"Optimal decoding" of neural responses entails identifying a simple transformation of the responses, $f(R)$, which will preserve as much of the MI as possible. A transformation that extracts all the information in the responses, so that

$$I(f(R), S) = I(R, S), \tag{3}$$

is known as a 'sufficient statistic' in classical statistical theory (Schervish, 1995, p. 113).

The data processing inequality holds when the real values of the MI are available. Instead, with real data, it is necessary to estimate the MI from the sampling distribution of the responses. MI estimates are, however, notorious for their bias (Treves and Panzeri, 1995; Panzeri and Treves, 1996; Victor, 2000). The naïve procedure computes MI as follows: a matrix of joint counts of stimuli and response statistics $n(f(r), s)$ collected from the data is used to estimate the joint probability of stimuli and responses $\hat{p}(f(r), s)$, using the maximum likelihood estimator $\hat{p}(f(r), s) = n(f(r), s)/n$. Then, the MI of the empirical distribution $\hat{p}$, is calculated by

$$I[\hat{p}(f(r), s)] = \sum_{x,y} \hat{p}(f(r), s) \log \frac{\hat{p}(f(r), s)}{\hat{p}(f(r))\hat{p}(s)}.$$

Finally, one has to estimate the MI of the 'true' underlying distribution $I[p(f(r), s)]$ from the empirical MI $I[\hat{p}(f(r), s)]$.

For finite samples, the most naïve MI estimator, taken to be the MI of the empirical distribution, is positively biased (Treves and Panzeri, 1995), i.e., on average, we have $I[\hat{p}(f(r), s)] > I[p(f(r), s)]$, in a manner that depends on the ratio between the number of bins in the count matrix and the number of measurements. For example, even if the stimuli do not modulate spike activity (so that $I[p(f(r), s)]$ is 0), the experimentally observed joint distribution will typically not be precisely separable into the product of the observed marginals $\hat{p}(x) = \sum_y \hat{p}(x, y)$ and $\hat{p}(y) = \sum_x \hat{p}(x, y)$, and so the estimated MI will be strictly positive. It has been shown (Treves and Panzeri, 1995; Victor, 2000; Paninski, 2003) that the bias is proportional to the number of occupied bins in the joint-distribution matrix $p(x, y)$. This is the number of possible stimulus-response combinations, and it may exceed the total number of measurements. Thus, the experimentally unavoidable severe under-sampling of the joint distribution matrices causes a large bias in information estimation.

The problem that we deal with in this paper is therefore two-fold. On the one hand, we are looking for a simple transformation $f(R)$ that will keep as much information about the stimuli as possible. On the other hand, in order to do that, it is necessary to compute MI of stimuli and responses in an extremely high-dimensional space, which will require ensuring that the bias does not result in overestimation of the MI. In order to solve this second problem, a number of estimation methods are used, and they are calibrated using simulated benchmark data that are consistent with much of the experimental data, produced using inhomogeneous Poisson models based on the experimental data (see Appendix I).

Reduced Statistics of Spike Trains

There are some natural candidates for the data reduction function f: these include the total spike count, the first spike latency, or the *mean response time*, defined as:

$$l = \frac{1}{n}\sum_{i=1}^{n} t_i, \tag{4}$$

where t_i are the times of the spikes following sound onset within some response window (64 ms in our case). The mean response time is identical to 1st spike latency when the response consists of a single spike. However, when the response has more spikes (as is often the case in the data analyzed here), it potentially encapsulates different information about the spike train than the 1st spike latency.

As we will show below, neither spike counts nor the mean response time by themselves are sufficient statistics for the data analyzed here. An example of a slightly more complex statistic is the 2-dimensional vector comprising spike count and mean response time for each response. We will show below that this statistic carries essentially as much information about the stimuli as the full responses.

Methods of MI Estimation

As discussed above, once the strategy of using a data-reduction function f is selected, the main problem for establishing the sufficiency of f is that of comparing the MI estimated from the full responses with the MI estimated from the reduced responses. The MI estimated from the full responses is especially hard to compute, because of the high dimensionality of the space of all possible spiking responses, which leads to large, under-sampled joint-distribution matrices and to large bias. There are a number of different approaches for solving this under-sampling problem. Naively, the simplest approach is to use coarser binning, both in stimulus space and in the response space. Binning is a data reduction step, and due to the data processing inequality, will result in lower MI of the binned data relative to the MI of the raw data. On the other hand, a smaller number of bins improves the estimates of the joint distribution $\hat{p}(x, y)$, and therefore reduces the bias. There is therefore an inherent tradeoff between the goal of preserving the raw MI, which requires small bins, and the goal of reducing bias, which requires large bins. This tradeoff is not unique to MI estimation, but is in fact inherent in various problems that involve density estimation (Duda et al., 2000).

A related approach is the use of decoding algorithms. A decoding algorithm generates a binning scheme on the response space, in that all responses that are assigned to a specific stimulus are binned together. Thus, the confusion matrix generated by applying the decoder to the data is in fact a coarser version of the full stimulus-response joint distribution. Decoding methods have been extensively used for studying neural codes in the past (Bialek et al., 1991; Rolls et al., 1997). Middlebrooks and his coworkers studied the coding of space in auditory cortex of cats (Furukawa and Middlebrooks, 2002), using decoding algorithms based on artificial neural networks. Victor and Purpura (1996) suggested the use of a metric based on the similarity of spike patterns and applied it to recordings in the visual cortex. Machens et al. (2003) used a different metric method to study the encoding of natural sounds by grasshopper auditory receptor neurons.

Recently, two fundamentally different approaches have been suggested for MI estimation. The first, by Panzeri and Schultz (2001) expands the MI in terms of moments of the joint distribution. The expansion is truncated after the 2nd order. This expansion has the great advantage that different information sources are separated explicitly. Its main usefulness has been in separating the contributions of firing rates and correlations to the MI for simultaneous recordings of a number of neurons (Panzeri et al., 1999a, 1999b, 2003; Petersen et al., 2001, 2002). However, the error involved in truncating the series is difficult to assess. In fact, in its applications, the information estimates derived by series expansion are usually compared with information estimates derived by the direct method (see below) in order to confirm their applicability.

In a different attempt to circumvent the issue of binning, Victor (2002) introduced a binless approach to MI calculations for continuous data. In this approach, estimates of the density $p(x, y)$ are derived from the 'distance' between a sample point and its 'neighbors'—small distances correspond to high density and large distances to low density. Unlike arbitrary binning, this approach allows the resolution to be determined in a data-dependent adaptive manner.

We applied four approaches to the data: (i) using joint distribution matrices directly with explicit bias-information loss tradeoff optimization (the adaptive direct estimates, AD); (ii) a decoding algorithm (due to Treves; Rolls et al., 1997, DEC); (iii) the 2nd order series expansion (Panzeri and Schultz, 2001, 2ND); and (vi) Victor's binless algorithm (Victor, 2002, BINL).

For all of these methods, except for Victor's binless algorithm (BINL), the temporal resolution must be selected. Larger time bins correspond to coarser representation of the data and therefore to smaller MI. On the other hand, larger time bins, by reducing the size of the response space, can also improve the stability of the estimates. To perform this tradeoff explicitly, we used a number of temporal resolutions, with time bins ranging from 1 to 32 ms, covering a fixed response window of 64 ms. At each resolution, the original spike train was "resampled" to yield a new train which had a 0 in each time bin that did not include any spikes and 1 in bins containing $\geq$1 spike. Each method was applied to these resampled spike trains and the maximal MI (after bias correction) over all temporal resolutions was used (see Reich et al., 2001, for a similar maximization procedure).

Methods (ii)–(iv) have been described previously in the literature. The full details of all estimation procedures are described in Appendix II. Here only the AD estimates (method (i)) will be described in more detail.

The AD estimates are computed using an explicit trade-off between MI reduction and bias reduction caused by coarse binning. We start with a matrix that is based on many bins with small probabilities. The naïve MI (including the bias) and a bias correction are computed for this matrix (methods for bias estimation are described further below). The matrix is then reduced, step-by-step, by joining together rows and columns (resulting in coarser binning), and the MI and bias are recomputed. The reduction continues until only a single row (or column) remains. The result is a set of decreasing MI values and a corresponding set of decreasing bias values. The MI is estimated by the largest difference between the two.

The matrix was reduced by iteratively merging the row or column with the smallest marginal probability (i.e. the lowest number of observations) with the neighboring row or column that had the smaller marginal probability. The neighborhood relationship was the natural one for count and latency data. For spike pattern data, the patterns were grouped according to spike counts and, within each iso-count group, ordered first by the time of the latest spike in the pattern, then by the time of the last-but-one spike and so on. Although other orderings are possible, it will be demonstrated below that this order gives rise to reasonable MI estimates and therefore other orderings were not tested.

Because in our data small marginal probabilities (at least in the initial binning steps) corresponded to rare response values, the routine initially tended to join rare response values into larger bins. The result was an inhomogeneous binning of the response range, in which different bins tended to have comparable marginal probability rather than comparable width.

This approach is similar to the method of sieves used by Paninski (2003) in his theoretical analysis of entropy and MI estimation from experimentally-measured count matrices. However, while in the method of sieves the coarser binning is selected a-priori, here it is generated adaptively in a way that depends on the empirically observed distributions. The main concern in selecting the binning adaptively is that the final estimate will be too large ('overfitting'). In our simulations, over-estimation was found to be negligible. The reason that our adaptive algorithm largely avoids over-estimation is probably the fact that each reduction step is based on marginal distributions and on the existence of a natural topology in stimulus and response spaces, but not on any feature of the joint distribution of stimuli and responses. It therefore searches within a relatively small and smooth hypothesis space, which is known to improve generalization ability. We also investigated alternative rebinning strategies whose target is to keep the mutual information as high as possible (e.g. using the Agglomerative Information Bottleneck algorithm, Slonim and Tishby, 2000), and these indeed resulted in a significant over-estimation of the MI. Another reason why over-estimation was avoided is probably the fact that although many response values were possible, most of them were highly improbable. For example, a large number of responses, mostly in the ferret data, consisted of a burst that was tightly locked to the onset of the stimulus. In this situation, although the space of all possible spike patterns or all possible response times was still very large, most of the distribution of possible spike patterns was concentrated on a small number of bins (the entropy of the spike pattern distribution was small). Therefore, the standard bias correction was too large (see Paninski, 2003 for an analysis of this situation), making all estimated values somewhat small. Again, under these circumstances the maximization inherent in the AD method did not result in over-estimation. These points may be amenable to more formal theoretical treatment.

Our adaptive procedure is also similar to techniques used for testing hypotheses about homogeneity and similarity of distributions (see e.g. DeGroot and Schervish, 2001, chapter 9), in which equal weight bins are created by binning the values of the cumulative distribution function into equal width bins. Reich et al. (2001) used a similar approach, combining response bins when data were rare, but they did not try to optimize their binning.

A crucial issue is the choice of the bias correction term. For count and spike latency measures, we used the simplest bias correction (Treves and Panzeri, 1995). Panzeri and Treves (1996) suggested a more sophisticated scheme for count matrices. The two estimates are substantially different when the count matrix contains a large number of zeros. However, in most cases, the optimal MI estimates occurred when the matrix size was small enough for the two bias-correction procedures to give similar corrections. When analyzing full spike pattern data, a different bias correction strategy was used, since the procedure described above resulted in a consistent, although small, under-estimation of the MI. This more sophisticated bias correction method is described in detail in Appendix II.

Validation of the MI Estimation Methods

To validate the MI estimation methods for the specific context in which we applied them, we created a parametric model that generated simulated data with statistical properties that imitated those of our experimental data. Because the model was analytically tractable, the MI estimates generated by the various methods on benchmark data could be compared to precisely known values. The purpose of using a model was not for calculating the MI between stimuli and the recorded spike trains, but rather for evaluating the validity of MI estimation techniques using relevant simulated data. The model we used is an inhomogeneous Poisson process (IHPP), which we found to agree well with most of our data (Appendix I). The variable rates of the IHPPs used in our simulations were chosen to correspond to rates observed in our data. For this purpose, 32 ms long response segments (11–42 ms after stimulus onset for the ferret data, 19–50 ms after stimulus onset for the cat data) for each neuron were smoothed slightly (Hamming window of length 4 ms in time and a triangular window of length 3 in the stimulus direction) to give the rate functions used in the simulations. The smoothing of the experimental data was required in order to reduce the MI of the model into the range of the bias-corrected MI values recovered from the data. We chose 32 ms windows since calculations of the true MI of the underlying distributions proved not to be computationally feasible for longer segments. For the IHPP we were then able to calculate the probability of observing any particular spike train in response to any stimulus, and hence to compute the 'true' MI according to Eq. (1). The probabilities of all possible response patterns with up to 6 spikes were computed individually, while probabilities of all spike responses with ≥ 7 spikes were collapsed into a single number. To study the effect of combining the patterns with large spike counts in this way, the MI was computed while considering in detail only patterns with up to 5 spikes, representing a coarser representation of the full joint distribution of stimuli and spike patterns. This decreased the MI by less than 0.01 bit (representing $<2\%$ of the MI) in all cases except for one single neuron in the cat dataset, where the decrease was 0.045 bits, representing about 12% of the MI.

Two types of simulations were performed. The first was a detailed study of the behavior of the MI estimates under conditions matching those of the experiments (40 repetitions for each of 24 stimuli for the ferret data, 20 repetitions for each of 15 stimuli for the cat data). Ten simulated samples were generated from each scenario, and the biases and variances of the estimated MI values were computed. The second set of simulations was a study of the asymptotic behavior of the MI estimates. For each simulated scenario, one sample of sizes 20, 40, 80, 160 and 320 repetitions per stimulus were generated and all estimates computed.

Decompositions of the MI

The mean response time is undefined in trials that do not have any spikes. This does not pose a problem for MI estimation: a special value is assigned to these trials, and the joint distribution of stimuli and responses includes the joint probability of this special value and the various stimuli. However, this means that the MI between stimuli and mean response time is a mixture of two sources: part of the MI is related to the probability of spike occurrence, and therefore related to spike counts, and the other one is purely timing information. It is possible to decompose the MI between stimuli and mean response time explicitly between the two information sources. Let l be the mean response time, and let z be the random variable whose value is 0 if there were no spikes and 1 if there was one or more spike in a trial. Let $l(z > 0)$ be the mean response time, except that the sampling space is limited now to those trials that had at least one spike. Then the following decomposition holds (see Appendix III for proof):

$$I(l; S) = I(l(z > 0); S)p(z > 0) + I(z; S), \qquad (5)$$

where the first term on the right hand side can be interpreted as the residual, 'pure', timing information, and the second term is the contribution of the differential distribution of null trials vs. trials that had at least one spike.

Although this decomposition was used explicitly for latency only, it can also be used for information between stimuli and spike counts. If c is the random variable whose value is the number of spikes in a trial, and $c(z > 0)$ is the random variable whose value is the number of spikes except that the sampling space is limited now to those trials that had at least one spike, the same decomposition holds:

$$I(c;S) = I(c(z>0);S)p(z>0) + I(z;S). \qquad (6)$$

Although less natural in the case of counts, this decomposition can be interpreted as separating the MI between stimuli and spike counts into a contribution of the residual distribution of counts across trials when there were spikes, and the contribution of the differential distribution of null trials vs. trials that had at least one spike. The advantage of this decomposition is the explicit factoring of one (in fact the main) source of redundancy between spike count information and mean response time information about the stimuli, namely $I(z;S)$.

Results

Overview

The current paper consists of two major parts: (1) Comparing and calibrating MI estimation methods using simulated data that approximate the actual recorded spike trains, and (2) estimating MI in recorded spike trains using optimal and simplified statistics, in order to find a reduced measure that can be used as the 'neural code'.

We started by a detailed comparative evaluation of the four estimation procedures, using a benchmark model for which the true MI could be calculated exactly. For the purpose of this benchmark, we show (in Appendix I) that an inhomogeneous Poisson process (IHPP) can be used to produce adequate simulated data. Evaluating the estimation methods in simulations revealed that bias due to sampling errors in the estimation of the underlying joint probability distributions is the dominant error term in all estimation methods. However, the four methods differed considerably in how much they were affected by such bias, and in how successfully the bias could be compensated for. We found that the adaptive-direct method (AD) performed considerably better than the others in this respect, and, after appropriate bias correction, it produced highly reliable estimates of the true MI of the simulated data.

We then studied the recorded spike trains, and compared information content extracted using different statistics. At this stage, the calculations were performed *without* assuming any specific distribution (such as IHPP) of the spikes. We show that the general features of the MI derived from the recorded spike trains using the different estimation methods were similar to those found in the IHPP simulations. We therefore used the estimates produced by the AD method as the best estimates of the MI in the recorded spike trains, and compared these to the MI carried in spike count or mean response time only. The MI carried by spike count alone varied from neuron to neuron, representing at times as little as 20% of the full spike pattern MI. In agreement with previous studies (e.g. Furukawa and Middlebrooks, 2002), we found that a timing measure, in our case the mean response time (Methods, Eq. (4)), is often more informative than spike count. Interestingly, we also found that, together, the reduced measures of spike count and mean response time captured essentially all the information about stimulus identity inherent in the spike train. In the remainder of this section we describe these key findings in detail.

Nature of the Data

The two datasets analyzed here were collected in primary auditory cortex (A1), but in different species, with different stimuli, and under different anesthetic conditions. We first describe general characteristics of each dataset.

In the ferret dataset (ferret A1 under barbiturate anesthesia), the majority of neurons (∼61%) responded to virtual acoustic space (VAS) stimuli with a short onset burst only. Figure 1A shows a representative example in raster plot format. In the example shown, both the mean response time and the mean spike count varied as a function of sound source position. An onset burst followed by between 1 and 3 later response peaks, with typical latencies between 120 to 400 ms, was also reasonably common (∼36% of all sampled units). Neural responses lacking a strong onset response were very rare (∼3%) and were not included in the information theoretical analysis. Similar phasic responses to VAS stimuli have been described in A1 of the cat (Brugge et al., 1996).

A representative example of a spatial response profile constructed from spike counts is shown in Fig. 1B. Since the five samples taken at each source direction were insufficient to provide the joint distribution estimates required for an information theoretical analysis, we pooled responses from sets of 8 adjacent virtual sound source directions. This yielded samples for 24 "sectors", each covering about $45^\circ \times 45^\circ$ of acoustic space (sector boundaries are outlined in Fig. 1B), with 40 responses sampled within each sector. The analysis

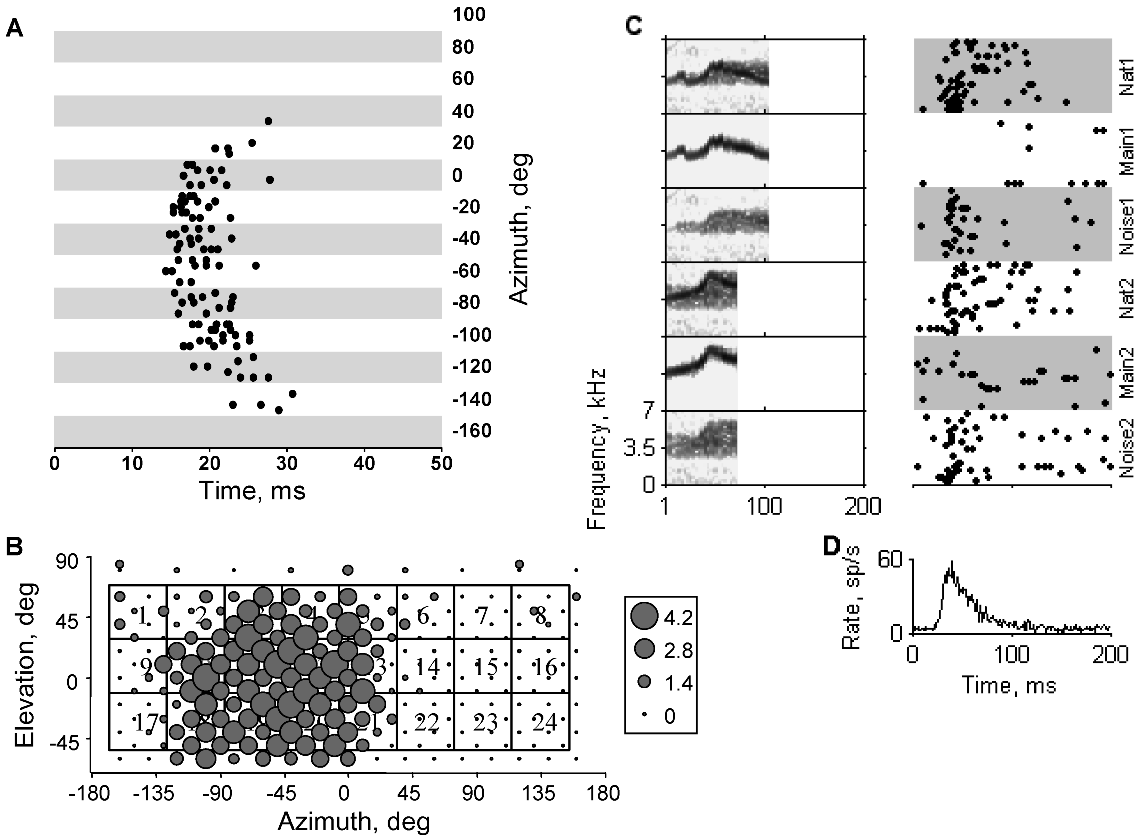

Figure 1. Responses of neurons in A1. A: Raster plot of individual responses to VAS stimuli for a single unit recorded in ferret A1. Each small black dot indicates the timing of one action potential. Each row of dots shows the response pattern to one stimulus presentation. Virtual source azimuths are indicated on the right. All responses shown in this plot were collected at 0° stimulus elevation. B: Spatial response field for the unit shown in A. Each gray circle shows the mean response to a VAS stimulus presented at the virtual source direction indicated by the axes. The diameter of the circles indicates the mean number of spikes evoked by one stimulus, according to the scale shown on the right. For the information theoretical analysis, responses to stimuli from adjacent directions were pooled according to the 24 sector boundaries shown as gray lines. C: Spectrograms of the stimuli (left) and raster plots of the responses of a cat neuron to those stimuli (right). Responses to two natural stimuli (Nat1 and Nat2) are presented, together with two versions of each: the cleaned main chirp (Main1 and Main2), and the difference between the natural sound and the main chirp (Noise1 and Noise2). D: Summed PSTH for the responses of this neuron to all 15 stimuli (see Methods). Note the long time scale of the decay of the response.

described here therefore effectively measures an upper limit of how reliably one can estimate from which of the 24 sectors a sound originated by observing a single response burst.

Figures 1C and D show some of the stimuli and the resulting responses of an A1 neuron in the dataset from halothane-anesthetized cats. The summed PSTH (Fig. 1D) shows much longer response durations than the responses in the barbiturate-anesthetized ferret (note the different scales of the abscissae in Figs. 1A and D). In fact, responses occurring ≥40 ms after stimulus onset were as common and as strong as those between 0 and 40 ms after stimulus onset (Bar-Yosef et al., 2002).

Figure 1C shows responses to three versions (Natural, Main and Noise, see Bar-Yosef et al., 2002) of two of the stimuli. Response patterns varied both within versions of the same stimulus and across stimuli in an idiosyncratic way (Bar-Yosef et al., 2002).

Response Characteristics

Figure 2 illustrates the relationships between the responses and the joint distribution matrices that are required in order to compute the MI. The two upper rows show data from two neurons in ferret A1, and the two lower rows show data from two neurons in cat A1.

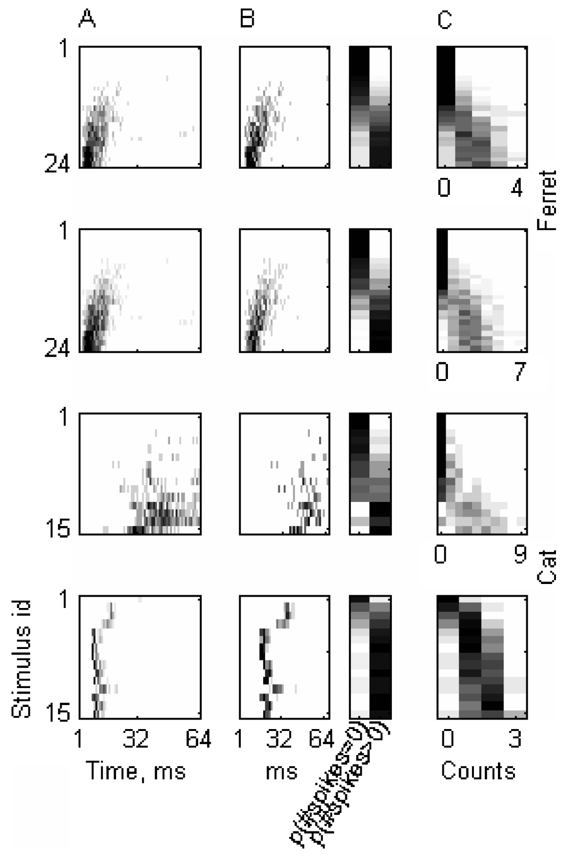

Figure 2. Responses of two neurons in ferret A1 to virtual space stimuli (two upper rows) and two neurons in cat A1 to bird chirps (two lower rows). A: Response planes. Instantaneous firing rate is plotted as a function of stimulus identity (id, on the ordinate) and time after stimulus onset (on the abscissa). Stimuli were ordered according to their mean spike counts. Gray scale saturation (top to bottom): 227, 332, 350, and 665 spikes/s. B: Joint distributions used for quantifying information in spike timing. The left panel represents the joint distribution of mean response time and stimulus id. When the responses are phasic (top two rows and bottom row), this histogram is similar to the response plane. Gray scale saturation: 1.5%, 1.7%, 2.8%, 3.4%. The right panel represents the joint distribution of the probability of having 0 spikes or >0 spikes and the stimulus id. Gray scale saturation: 4%, 4%, 6.6%, 6.6%. C: Joint distributions of spike counts and stimulus id. Gray scale saturation: 2.9%, 2.9%, 4.6%, 4.3%.

Column A consists of response planes, i.e. plots of the instantaneous firing rate as a function of time after stimulus onset (on the abscissa) and stimulus identity (on the ordinate). Thus, each row displays the PSTH in response to one stimulus. The stimuli are ordered by increasing values of average spike counts, for each neuron separately. Thus, for each neuron, stimulus no. 1 evoked the least number of spikes, and stimulus no. 24 (ferret) or 15 (cat) evoked the largest number of spikes.

Column B shows the joint distributions used for quantifying the information in spike timing. The timing of spikes in each trial was summarized by a relatively coarse latency measure, the mean response time over all spikes in the burst (Eq. (4)). Latency is undefined for responses that contained no spikes. Therefore, we quantified the joint distribution of mean response time and stimulus by two histograms: the first is the histogram of the actual values of the mean response times computed for trials that had at least one spike (left histogram in Fig. 2B). In this histogram, the abscissa represents the mean response time and the ordinate represents the stimulus identity. The histogram is normalized, so that the gray levels represent the joint probability of observing a given mean response time and a given stimulus. The second histogram (right histogram in Fig. 2B) represents the probability of having at least one spike in the response. In order to be a true probability distribution, it also needs to encode the probability of the complementary event of having no spikes, and the sum of these two is equal to the probability of the presentation of the stimulus. Both of these histograms carry information about the stimulus, and their contributions were combined when computing the MI between mean response time and stimulus (see Methods for details).

Figure 2C displays the joint distribution of spike counts and stimuli. The spikes were counted in a window of 64 ms starting at stimulus onset. This window was long enough to include the most informative parts of the responses in both datasets, although particularly in the cat dataset some responses extended beyond the 64 ms window.

The two neurons from ferret A1 had responses that consisted of short bursts of action potentials, whose magnitude and latency depended on stimulus direction. Thus, mean response time and spike counts were highly correlated, with shorter mean response times associated with higher spike counts (compare joint distribution matrices in Figs. 2B and C).

The two neurons from cat A1 (two lower rows) showed longer latencies than the neurons from ferret A1. To some extent, these longer-latency responses are due to the relatively slow rise times of the natural stimuli used here (Bar-Yosef et al., 2002), but may also result from the different anesthetics used in the two datasets (Rotman et al., 2001; Bar-Yosef et al.,

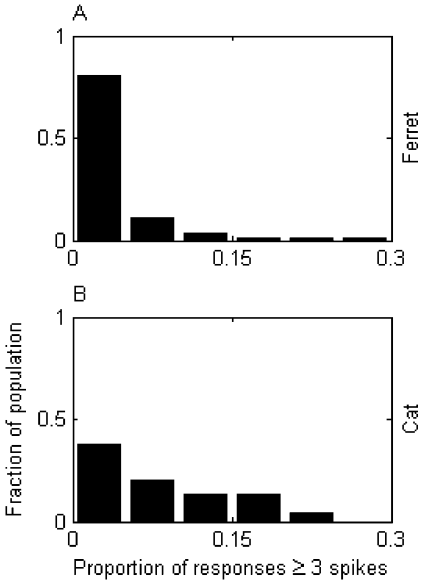

Figure 3. Distribution of large responses in the population. The proportion of single responses that contained ≥3 spikes/stimulus is plotted on the abscissa, while the ordinate shows the proportion of neurons with these responses. A: Ferret dataset. B: Cat dataset.

2002). The correlation between spike counts and mean response time was much weaker, if present at all, in these neurons.

The total spike count in the 64 ms window following each stimulus onset was relatively low, as commonly observed even for highly efficient stimuli in auditory cortex. This can have strong implications for the nature of the neural code. For example, if the responses have at most a single spike, the spike time (with a special value encoding the absence of spikes) is a sufficient code. Figure 3 shows the probability distribution of there being three or more spikes in a single response over the population. In the ferret dataset (Fig. 3A), most responses consisted of less than 3 "onset" spikes, but 8% of the neurons had a probability >0.1 of having ≥3 spikes in a single response. In the cat dataset (Fig. 3B), spike counts were larger—over 42% of the neurons had a probability >0.1 of responding with ≥3 spikes per stimulus, and 11% had a probability of ≥0.3 of having ≥2 spikes in a single response. Thus, despite generally low firing rates, there were numerous instances where the response comprised ≥3 spike, indicating that the number of spikes (and not only the occurrence of a burst) could contribute to the neural code.

Comparing MI Estimators: Simulation Results

To compare the reliability of different MI estimators, we created a model that generates simulated data with the same characteristics of the real data (see Methods). Because the information content in the simulated data is known exactly, we can use these simulations to determine the accuracy of MI estimates produced by various methods. Simulations are therefore used for validating the MI *estimation methods*, rather than obtaining MI *estimates*. Figure 4A compares MI estimates from the four different MI estimators (described in Methods and in Appendix II) with the true MI values. The mean MI for each estimator is plotted against the true MI from the same simulated scenario (for spike patterns of up to 6 spikes per simulated response). The number of repeats per stimulus in each simulated dataset matched those of the electrophysiological data, 40 for the ferret dataset and 20 for the cat dataset.

Victor's binless estimates (BINL, see Methods for details of this and the other estimation methods) and the estimates based on Panzeri's 2nd order expansion (2ND) show a consistent negative bias. In addition, the BINL estimates have a relatively high variance. In contrast, the AD estimates exhibit very little bias, except for the simulations based on the responses of one neuron from the cat dataset. This is the same neuron whose true MI was underestimated due to the limitation of the calculation to patterns of up to 6 spikes (see Methods), and is the neuron with the highest spike counts in the dataset, with a high probability of having 6 and 7 spikes within the counting window. Finally, the estimates based on Gaussian decoding (DEC) show a small positive bias at low MI values, but are essentially without bias at MI values above 0.5 bits.

The bias and variance of these estimates were reduced when the simulated dataset was increased to 320 trials per stimulus. This is expected theoretically since all the methods discussed here produce consistent estimators in the asymptotic limit of infinite data. Figure 4B shows the various MI estimates with the number of trials per stimulus used in the experiment (40 for the ferret data, 20 for the cat data) and 320 trials per stimulus (triangles). The bias is clearly reduced in the BINL, 2ND and AD estimates. The DEC estimate also shows reduced scatter at 320 trials per stimulus, but for most points it seriously underestimated the MI, whereas for some points it produced large

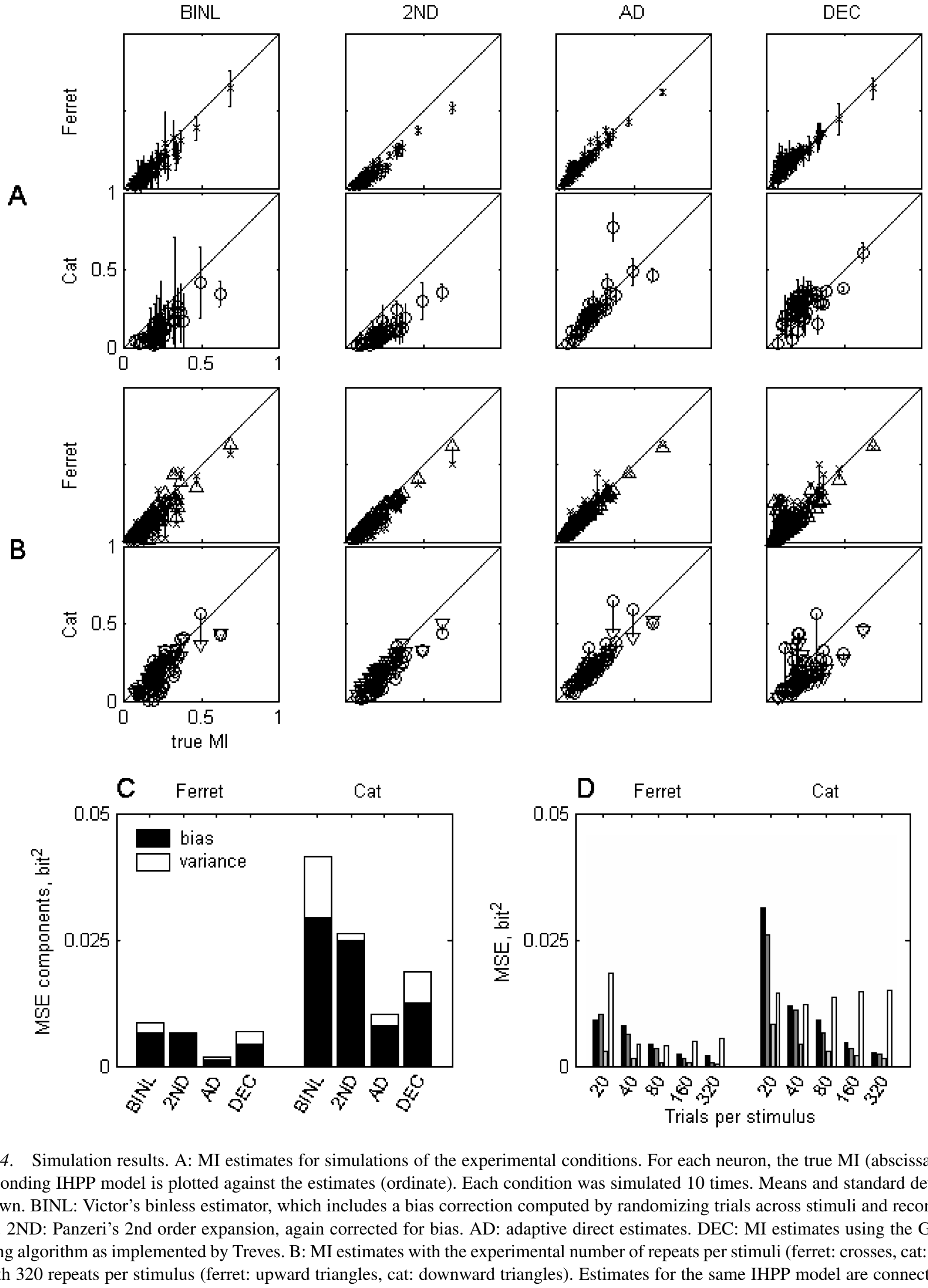

Figure 4. Simulation results. A: MI estimates for simulations of the experimental conditions. For each neuron, the true MI (abscissa) of the corresponding IHPP model is plotted against the estimates (ordinate). Each condition was simulated 10 times. Means and standard deviations are shown. BINL: Victor's binless estimator, which includes a bias correction computed by randomizing trials across stimuli and recomputing the MI. 2ND: Panzeri's 2nd order expansion, again corrected for bias. AD: adaptive direct estimates. DEC: MI estimates using the Gaussian decoding algorithm as implemented by Treves. B: MI estimates with the experimental number of repeats per stimuli (ferret: crosses, cat: circles) and with 320 repeats per stimulus (ferret: upward triangles, cat: downward triangles). Estimates for the same IHPP model are connected with lines. Each symbol is the result of a single simulation. Columns as above. C: Summary of the mean square error (MSE) for the simulations under the conditions of the real datasets. The black bars are the contribution of the bias term to the MSE, while the white bars indicate the additional contribution of the variance. D: MSE for the asymptotic results. In each group of bars, the order of the methods (from dark to light gray) is BINL, 2ND, AD and DEC.

over-estimates of the true MI. This instability resulted from the maximization over different temporal resolutions. When only finer temporal resolutions were considered for maximizing the DEC estimates, this instability was reduced (data not shown).

The simulation results are summarized in Figs. 4C and D. The mean squared error (MSE) of the four estimates is displayed in Fig. 4C for the number of trials per stimulus used in the experiments. Since each condition was simulated 10 times, it was possible to partition the MSE into a bias (black) and a variance term (white). These data demonstrate that the main limitation of all the estimators is bias, rather than variance: the bias term contributes more than half the MSE in all cases, and over 90% in some cases. The 2ND and the AD estimates exhibited the lowest variance. However, the 2ND estimates had a much larger bias than the AD estimates. The dominance of bias over variance in MI estimation has been demonstrated rigorously by Paninski (2003) for MI estimation from count matrices. We used different estimation methods, but nevertheless also found that bias is the dominant error term, suggesting that it is the major source of error in MI estimation methods in general.

Figure 4D illustrates the decrease in the MSE as the number of trials is increased. Since only one simulation was performed for each combination of scenario and number of trials per stimulus, the MSE could not be partitioned into bias and variance. For unbiased estimators, the MSE is expected to decrease by a factor of two for each doubling of the number of trials. In most cases, the reduction in MSE was slower than expected, probably due to the presence of an important bias term, even when using 320 trials per stimulus.

The data presented in Fig. 4 have two implications. First, bias is the main source of error in MI estimates for all methods used here. Second, the AD estimates had the smallest MSE, and generally produced highly accurate estimates of the information content of the simulated data, despite the difficulties inherent in estimating MI from large, sparsely sampled stimulus-response matrices.

Comparing MI Estimators: Real Data Spike Patterns

Figure 5 compares the four estimates of the full spike pattern MI for the experimentally recorded spike trains. The scatter plots show the AD estimates against each of the three other estimates (Fig. 5A: BINL; Fig. 5B: 2ND; Fig. 5C: DEC). The BINL and AD estimates (Fig. 5A) are widely scattered along the diagonal, which is probably largely attributable to the high variance of the BINL estimator revealed by the simulations (Fig. 4). In addition, the BINL estimates are consistently smaller than the AD estimates. The 2ND estimates (Fig. 5B) are more tightly correlated with, although still consistently smaller than, the AD estimates for both datasets. Finally, the DEC estimates tended to be larger than the AD estimates for both datasets (Fig. 5C). These features are all consistent with the simulation results (Fig. 4A). Thus, the pattern of results for the real and simulated data matches reasonably well, assuming that the AD estimates are indeed close to the (unavailable) true MI.

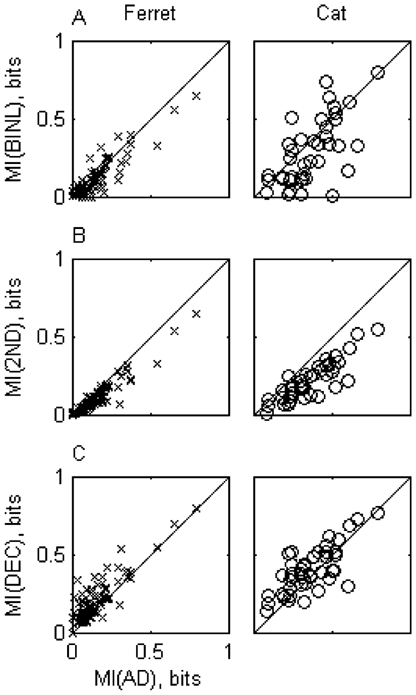

Figure 5. Estimates of full spike pattern MI from the experimental data. The AD estimates are displayed on the abscissa, and the ordinate shows each one of the other three estimates. A: BINL. B: 2ND. C: DEC.

The MI estimates for the real data are generally fairly small: for example, the AD estimates for the ferret dataset range from 0 to 0.79 bits/trial, with a median of 0.12 bits/trial. For the cat dataset, they range from 0.08 to 1.4 bits/trial, with a median of 0.33 bits/trial. Both distributions are highly skewed: for example, in the cat dataset, only two neurons had MI values greater than 0.8 bits/trial. The MIs in the cat dataset are significantly larger than in the ferret dataset. However,

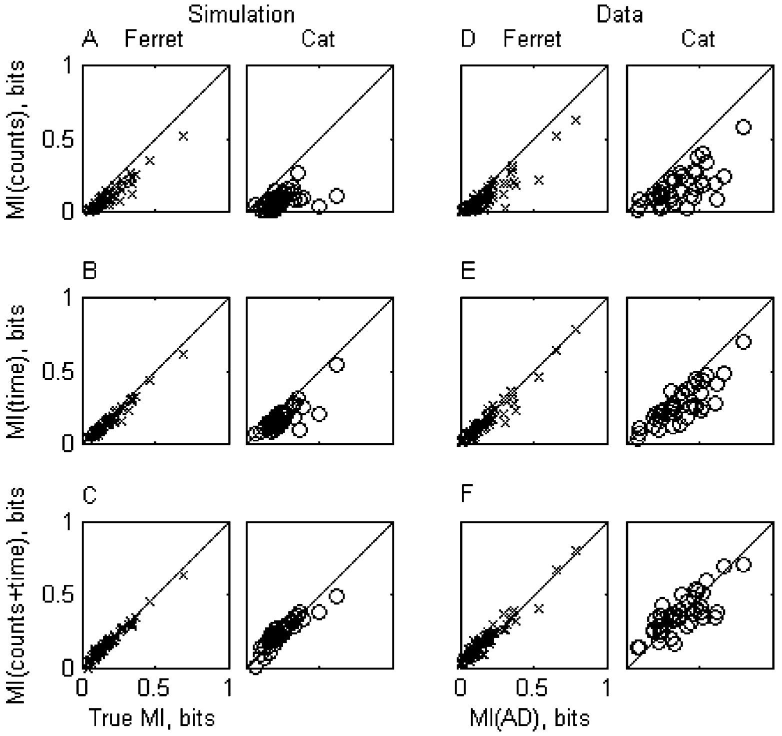

Figure 6. MI derived from the reduced measures. A–C: Simulation results: true MI (on the abscissa) against the reduced measures. A: MI between counts and stimuli. B: MI between mean response time and stimuli. C: MI between (counts, mean response time) and stimuli, computed from the simulated spike trains. D–F: MI estimates derived from the experimental data: AD estimate (on the abscissa) against the reduced measures. D: MI between counts and stimuli. E: MI between mean response time and stimuli, computed from histograms. F: MI between (counts, mean response time) and stimuli, computed using histograms.

because the spike counts in the cat data are, on average, also ~2.5 times larger than in the ferret data, the mean MI per spike is essentially the same in each case: 0.43±0.33 (mean ± SD) bits/spike in the ferret dataset, 0.43 ± 0.20 bits/spike in the cat dataset.

The ferret MI values are also substantially smaller than those reported by Middlebrooks and colleagues for neurons in cat auditory cortex using free-field stimulation from multiple directions in space, where values close to 1 bit were commonly found (e.g. Furukawa and Middlebrooks, 2002). However, the decoders used by these authors operated on superpositions of 8 individual trials. Consequently, they do not reflect the MI for a single neural response, but rather the MI for the pooled "ensemble response" of 8 independent copies of the same neuron. We expect that if the decoders used by Middlebrooks and colleagues operated on single trial responses, their observed MI values would be comparable in magnitude to those presented here.

MI of Spike Counts and Burst Latencies: Simulations

Figure 6A–C shows simulation results for the MI between stimuli and the reduced measures, spike counts and mean response time. Because the simulations were produced using a known probability model, the joint distributions of the stimuli and the reduced measures could again be computed exactly within the approximation that only spike patterns comprising up to 6 spikes are considered individually.

In Fig. 6A, the MI between spike counts and responses is compared with the MI estimated from the full spike patterns. For the IHPP models derived from the ferret data, using spike counts only instead of the full spike patterns reduced the MI by about 20%. In contrast, in the IHPP models derived from the cat data, using spike counts instead of full spike patterns reduced the MI more severely, with some units losing up to 90% of the MI. The MI between stimuli and mean response time is shown for the simulations in Fig. 6B. The picture is similar to that for the spike counts, except that the information loss is less pronounced.

Finally, Fig. 6C shows the mean estimated MI between the pair (spike counts, mean response time) and the stimuli where simulated spike trains were used to generate the joint-distribution matrix. In this case, the points cluster tightly around the diagonal, showing that, for the IHPP models estimated from the data, essentially all the information available in the spike trains about the stimulus can be extracted from this reduced 2-dimensional statistic.

MI of Spike Counts and Burst Latencies: Real Data

The MI between stimuli and the reduced measures of the real data is displayed in Figs. 6D–F. Figure 6D plots the spike count MI against the MI derived from the full spike patterns using the AD method. Reducing the spike patterns to spike counts resulted in a significant loss of MI. The amount of information lost is highly variable, reaching in some cases over 80% of the estimated full MI. The information loss is particularly pronounced in the cat dataset. Figure 6E compares the full MI and the MI derived from the mean response time. For both datasets, there is still a significant decrease in information, although less than for spike counts alone.

Figure 6F shows the results for the MI between the 2-dimensional variable (spike counts + mean response time) and the stimuli. These estimates cluster around the diagonal, showing that the full spike pattern MI, as estimated by the AD method, and the MI in the joint distribution of count and mean response time are essentially equivalent. Indeed, the differences between the means of the AD estimates and of the 2-dimensional reduced measure were small: −0.01 bits for the ferret dataset and 0.03 bits for the cat dataset. This difference was not significant for the cat data (paired t-test, $t = 1.4$, df = 44, n.s.) but was significant for the ferret data (paired t-test, $t = -3.8$, df = 96, $p < 0.01$), probably because of the larger sample size. The correlation coefficients between these MI estimates were 0.96 for the ferret dataset and 0.83 for the cat dataset. The greater scatter in the cat data may be due to the smaller number of repeats per stimulus.

In summary, we conclude that in the two datasets analyzed here, spike count alone may fall a long way short of capturing all the information carried in a neural response train, whereas mean response time tends to transmit more information than spike count but still falls short of capturing all the available information. However, spike count and mean response time together carry essentially all the information about the stimuli available in the full spike patterns.

The information contained in the mean response time and in the spike counts may not be independent. In principle, the spike counts and mean response time could be redundant (their individual MIs add up to more than the MI of the joint count-latency measure), independent (their sum approximately equals the joint count-latency MI) or synergistic (their sum is less than the joint count-latency MI). We calculated a redundancy/synergy (RS) index for the two reduced measures of count and latency by subtracting the sum of the spike count MI and the response latency MI from the joint count-latency MI value:

$$\begin{aligned} RS = {} & I(counts, latency; stimulus) \\ & - I(counts; stimulus) - I(latency; stimulus) \end{aligned}$$

RS is negative for redundancy, zero for independence and positive for synergy between the individual reduced measures.

The RS histograms (data not shown) are unimodal. Their medians (−0.035 and −0.044 for the ferret and cat datasets respectively) are significantly smaller than 0 (Wilcoxon signed rank test, $p < 0.01$ for both), suggesting that, over the neural population as a whole, redundancy between count and timing information is the norm.

To study the redundancy between spike count information and mean response latency information, we decomposed the MI (Methods, Eqs. (5) and (6)) into a term related to the presence or absence of spikes in a trial, and a term that measures the information in those trials in which spikes actually occurred. Since the first term is the same for both spike counts and mean response time, it is potentially the source of the redundancy between the two sources of information.

Figure 7A plots the MI between stimuli and spike counts against the MI between stimuli and mean response latency. As expected, the spike count information is consistently smaller than the mean response time information. However, the two are highly correlated (ferret: $r = 0.96$, df = 95, $p \ll 0.01$; cat: $r = 0.74$, df = 43, $p \ll 0.01$). Furthermore, the scatter plots do not show any grouping of the data. Thus, the way stimulus information is coded in these two measures appears to be homogeneous across the population.

The correlation between the two information sources could be the result of the common term related to the

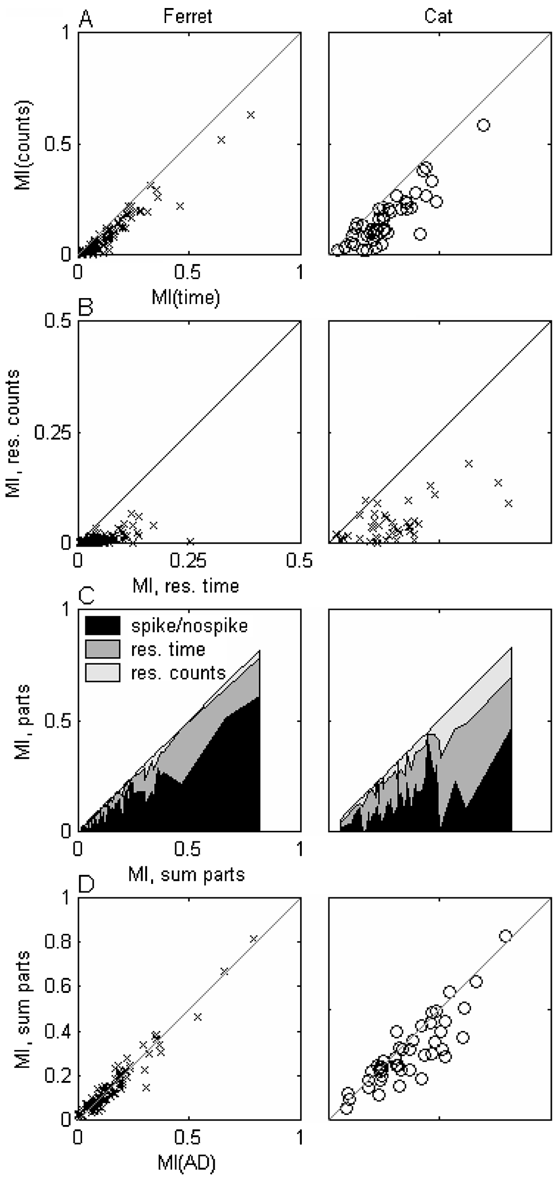

Figure 7. Decomposing the information. A: scatter plots of the full information in counts against the full information in mean response time. B: scatter plots of the residual information in counts against the residual information in mean response time (as in Eqs. (5) and (6), Methods). C: Information components for all neurons. The abscissa represents the sum of the three information components (presence/absence of spikes, residual spike count information, and residual mean response time information). The ordinate shows the stacked individual components. D: The sum of the three components plotted against the full MI.

presence or absence of spikes in a trial. As expected, once the common term is removed, the residual information in the mean response time is significantly larger than the residual information in the spike counts (ferret: $t = 12.5$, df $= 96$, $p \ll 0.01$; cat: $t = 5.96$, df $= 44$, $p \ll 0.01$). Furthermore, the residual information in counts and the residual information in mean response time show substantially lower correlation than when the full information is compared, although this correlation is still significant (ferret: $r = 0.52$, df $= 95$, $p \ll 0.01$; cat: $r = 0.47$, df $= 43$, $p \ll 0.01$). The scatter plot of the two (Fig. 7B) is still homogeneous, suggesting a continuum of coding properties in the cortical population.

Figure 7C displays the three contributions to the MI in the two datasets, plotted as a function of their sum. In both datasets, the largest contribution on average arose from the common term related to the presence or absence of spikes in a trial, followed by the residual information in mean response time, and finally by the residual information in spike counts. The relative contribution of each was, however, different in the two datasets. In the ferret, the average size of the three contributions was 54% (presence or absence of spikes), 37% (residual information in mean response time) and 7% (residual information in spike counts) of the full information, accounting in total for 98% of the full information in spike patterns (Fig. 7D). In the cat the average size of the three contributions was 31%, 40% and 11% respectively, accounting in total for 82% of the full information between stimuli and responses (Fig. 7D). The larger contribution of residual terms in those trials that had spikes in the cat dataset is probably accounted for by the larger number of spikes per response and by the richer temporal structure of these responses (Figs. 1–3).

The fact that all three information sources together essentially account for the full information in the ferret dataset, but only for about 80% of the information in the cat dataset, suggest that in the cat dataset there is a synergy between the counts and the latency information, at least in those trials that had spikes.

Discussion

Why Estimate the Full MI?

The MI between stimuli and responses is hard to compute for full spike patterns because of the high dimensionality of the response space. This is one of the reasons why spike counts are used extensively in neuroscience—they are a simple, and often reasonably informative, summary statistic of neurophysiological data. However, this study, like other previous studies (e.g. Rieke et al., 1997; Panzeri et al., 2001; Furukawa and Middlebrooks, 2002), clearly shows that there is much more information in neural responses than captured by spike counts alone. As discussed above, the MI between stimuli and spike patterns is an upper bound on the information that could be extracted from spike trains by any decoder or reduced response measure. This makes the MI between stimuli and spike patterns an absolute reference value against which any putative neural coding schemes should be compared if the neural code is to be discussed in a rigorous, quantitative manner.

There have been two arguments against the usefulness of the full MI estimates. The first argument is methodological: the full MI is hard to estimate accurately. The second is biological: even if it is possible to estimate the full MI, it is doubtful whether any biological decoder could extract anything like the full information. Our results indicate that, at least in the limited context tested here, neither argument applies.

Using the IHPP models we were able to show that it is possible to estimate the full MI quite accurately, even with relatively limited datasets. Although we cannot be absolutely certain that we estimated the full MI of the real data (since the exact underlying probability distributions are unknown), the very good performance of the AD method on the simulated data, together with the fact that the data are quite well approximated by an IHPP (see Appendix I), strongly suggest that our estimates of the full MI for the real data are valid. The success of the AD method is probably due to the fact that the responses are relatively sparse, rarely containing >4 spikes, and that most of the probability in the spike pattern space is carried by a set that is substantially smaller than all possible patterns. In information-theoretic terms, the entropy of the spike patterns is low.

Given the data processing inequality, one may doubt whether any physiological decoder can perform at values approaching the full MI. However, for the data analyzed here, the best estimates of the full MI can be closely approximated by simple reduced measures. The full MI is therefore an achievable upper bound, against which putative physiological decoders should be measured.

Generalizability

The main result of this paper is the demonstration that for the two datasets analyzed here, essentially all the information between stimuli and responses is captured by a pair of reduced measures, spike counts and the mean response time. In this sense, this pair of measures is a sufficient statistic—there is no measurable loss of MI when any features of the full spike pattern beyond the total spike count and the mean response time are ignored. This finding does not imply that other timing information (e.g. mean interspike interval, standard deviations of spike time or of interspike intervals) is not informative. But, once the spike count and mean response time are known, any of these other measures will add very little information about the stimuli.

How general are these conclusions? Our results certainly do not imply that the two statistics are joint sufficient in general. In other systems and for other stimuli their sufficiency will have to be checked again. Also, sufficient statistics do not have to be unique, and other simple statistics could be sufficient as well. However, some general features of the data analyzed here, which are commonly found in other systems, seem to be responsible for the sufficiency of relatively simple response statistics. The temporal patterns in the responses analyzed here are well suited for such a simple code. Indeed, most responses in our datasets appear as short bursts of spikes, whose timing depends on the stimulus. In the ferret dataset, spatial position modulated the strength and timing of the response burst, often in a correlated fashion (stronger bursts were associated with shorter latencies). The cat responses also often comprised phasic components, whose timing was modulated by the stimulus, although both mean spike counts and timing depended on the stimulus in complex ways (Bar-Yosef et al., 2002), and the two were less well correlated. Spike timing did not seem to be locked to specific acoustic features in the stimuli other than stimulus onset (Schnupp et al., 2001; Bar-Yosef et al., 2002). Thus, the timing differences used here to encode stimulus location or identity are probably generated by internal processing mechanisms, rather than being imposed by the physical structure of the sounds.

In the somatosensory system, where phasic responses are also very common, timing has been shown to carry a large amount of information about stimuli. Thus, even in the periphery, first spike latencies of populations of primary afferents carry information about the shape of objects during natural object manipulation (Johansson and Birznieks, 2004). In this case, spike timing is probably related to the sequence of activation of somatosensory receptive fields due to the active manipulation itself. This is similar to the use of timing for carrying information about sounds in phase-locked neurons of the auditory system. However, in somatosensory cortex, where timing carries important stimulus information (Panzeri et al., 2001; Petersen et al., 2002), the differences in spike timing do not seem to be directly related to external events. Thus, somatosensory and auditory cortices seem to use similar mechanisms for encoding stimulus identity by spike timing. A similar use of first spike timing has been suggested in the visual system. The speed of image categorization, measured psychophysically and with evoked potentials, and modeling studies both suggest that first spike latencies across populations of visual neurons are sufficient for encoding the identity of complex scenes (Van Rullen and Thorpe, 2002). In all of these cases, the structure of the responses suggests that a simple timing measure should carry a substantial amount of stimulus information.

Implications for the Neural Code

The finding that essentially all the information in the spike trains can be encapsulated by just two measures (spike count and mean response time) has two important implications, even within the limitations discussed above. First, in terms of data analysis, this result supports the use of reduced measures in analyzing stimulus/response associations. However, spike count alone falls well short of capturing the full information. Indeed, for some neurons in the cat dataset, this measure transmitted less than 20% of the available information. Thus, our results strongly argue for the routine use of additional, temporal measures when stimulus/response relationships are analyzed. On the other hand, very detailed or complex temporal measures may not be required. Simple measures such as first spike latency (useful when spontaneous activity is low), or the mean response time as defined here, may be sufficient.

The second important implication is that the counts and burst latencies could serve as the neural code, essentially without information loss. The decoding mechanisms used by downstream neurons therefore do not necessarily have to precisely time each and every spike. Instead, it is sufficient to have a mechanism that simply counts the number of spikes in the response, and marks their mean time. The feasibility of a decoding scheme based on spike latency has been discussed by others (Brugge et al., 2001; Furukawa and Middlebrooks, 2002; Jenison and Reale, 2003; Reale et al., 2003).

The joint sufficiency of spike counts and burst latencies as a neural code held under two very different experimental conditions, using different species (ferret and cat), different anesthetics (barbiturate and halothane) and very different stimulus classes (short, synthetic noise bursts delivered in VAS versus temporally complex natural stimuli with strong tonal components). Although the responses and their dependence on the stimuli are very different under these circumstances, the cortex appears to use the same very simple neural code in each case.

Appendix I: Testing the Data Against the Inhomogeneous Poisson Assumption

In order to demonstrate that a process is an inhomogeneous Poisson (IHPP), it is necessary to show that temporal correlations of all order are null. The IHPP has inherent limitations in its ability to account for neural data. For example, it does not have a refractory period. For a process with refractory period, the correlations between time bins that are less than the refractory period apart should be negative, which is impossible in an IHPP. However, under the appropriate conditions, such as the low spike counts that were encountered here, even refractory effects can give rise to only very small correlations. Thus, IHPP can approximate neural data quite accurately when applied correctly.

Checking temporal correlations of all orders is impossible to do with experimental data. Thus, we only checked first order spike distributions and second order correlations between spike counts in different time bins. In an IHPP, the spike distributions are Poisson while the correlations are null.

First, the expected distribution of spike counts, given the response rates for each stimulus, was computed under the Poisson assumption by simply summing the average spike numbers in each bin of the PSTHs. These expected count distributions were compared with the observed ones using a χ^2 test. To increase the power of the test, the χ^2 values for all stimuli were added up. Only 9/97 (10%) of the ferret neurons and 5/45 (11%) of the cat neurons had χ^2 values larger than the 10% critical value. Such deviations are expected from sampling noise. It can be concluded that, for 90% or more of the neurons, the spike count distributions can be deduced from the average rate using the Poisson assumption.

To check the size of the noise correlations between different bins, we used the values calculated in the 2nd order expansion procedure (Panzeri and Schultz, 2001). In the 2nd order expansion, the 2nd and 3rd components of the 2nd order contribution (Appendix II) are zero for IHPP. The 2nd component is generally non-zero when there are stimulus-independent correlations between spikes at different time points, and the 3rd component is generally non-zero when there are stimulus-dependent correlations between spikes at different time points. While individual correlations may be large because the normalization becomes unstable at high temporal resolution, the contribution of the 2nd and 3rd components of the 2nd order contribution to the MI can serve as a measure of the importance of non-Poisson correlations in the data to the MI calculations. Since the 2nd order expansion was computed for a number of temporal bin sizes (Appendix II), we used the components at the temporal resolution that gave the largest MI, which was usually 2–8 ms. The 2nd and 3rd components of the 2nd order term were larger than 20% of the size of the 1st order contribution in 2/97 and 7/97 neurons of the virtual space dataset, respectively, and in 1/45 and 6/45 of the neurons in the cat dataset. The contribution of higher order non-Poisson correlations to the MI was therefore generally small.

Furthermore, these values are not only low, they also represent an overestimate of the importance of the non-Poisson contributions, since they are biased (for example, the 3rd component is always positive). In the IHPP simulations, these components also had non-zero values, which are therefore rough estimates of the expected estimation bias. The experimentally-derived estimates were very close to these approximate bias values (all, except one estimate in the virtual space set and one estimate in the cat set, were within 10% of the estimated bias).

In conclusion, the data analyzed here can be represented adequately, up to sampling error, by an IHPP, because the spike distributions are consistent with the Poisson distribution and because the contribution of the non-Poisson correlations to the MI is negligible.

Appendix II: Details of the MI Estimation Procedures

(i) *The adaptive direct method.* The basic procedure of the adaptive direct method has been described in the Methods. The main problem in this procedure was related to the bias correction scheme applied at each step of the reduction process. When applying the standard Panzeri-Treves bias correction to simulations of spike trains, a small negative bias was observed. This bias persisted even with 320 trials per stimulus. The underlying cause is probably that the bias correction procedure tries to estimate the number of degrees of freedom in the histogram, which is essentially equal (for unsmoothed histograms) to the number of bins that could be non-zero (Panzeri and Treves, 1996). However, joint distributions of spike patterns and stimuli are very sparse: for example, for neurons with low spontaneous rates and ineffective stimuli, essentially only a few spike patterns (the 0 spikes pattern and 1-spike patterns) are expected to occur.

A similar problem occurs with other patterns. For example, suppose that a neuron responds to stimuli with a single spike whose latency is tightly linked to the stimulus. The set of possible spike patterns will consist of 1-spike patterns at various possible latencies, but for each stimulus, only a small number of these possible latencies may actually be observed. Thus, even with a large number of observations, the observed number of stimulus-response bins is much smaller than the total number of bins in the distribution matrix of stimuli and spike patterns. Formally, for the neurons described here, the joint distribution of stimuli and spike patterns is often supported by the boundary of the simplex of all possible joint distributions of stimuli and spike patterns (Paninski, 2003).

To address this problem, we modified the standard bias correction in the reduction procedure. Following Panzeri and Treves (1996), we estimated the degrees of freedom separately for each stimulus, then summed them together. However, the estimates suggested by Panzeri and Treves (1996) still resulted in an underestimate of the MI. Therefore, we estimated the relevant number of bins differently. For each stimulus s, the largest number of evoked spikes $m(s)$ was determined, and the estimated number of relevant bins for each stimulus was computed as

$$RB(s) = \sum_{n=0}^{m(s)} \frac{\#(n)}{n+\alpha}, \qquad \text{(A1)}$$

where $\#(n)$ is the total number of patterns with at most n spikes in the full data, and α is a parameter that depends on how many of these patterns we believe will be effectively populated in our data. The best value of α, as estimated in the simulations described below, was 0.6. When a matrix reduction step occurs, the number of spikes assigned to a joined bin is the maximum of the numbers assigned to each of the two original bins.

Intuitively, this bias correction term is based on the following considerations. The relevant bins should correspond to those spike patterns that have a substantial probability of being observed. The number of relevant bins for stimulus s estimated by the Bayesian Panzeri-Treves approach (1996) is $\sum_{n=0}^{m(s)} \#n(s) + \textit{correction}$ where $\#n(s)$ is the observed number of n-spike patterns in response to stimulus s, and the correction is related to the possibility that bins with non-zero probability are nevertheless empty because of sampling limitations. The correction term they developed is, however, based on assumptions that are not appropriate for spike pattern matrices, and results in incorrect bias estimates. Therefore, instead of estimating the number of spike patterns that could appear for each stimulus separately, all are counted (in fact, for n-spike patterns, the $\#n$ in Eq. (A1) is even larger—it is the total number of spike patterns with n or less spikes) and then this total number is divided by $n + \alpha$, as a way of distributing these patterns among stimuli. Thus, for example, any pattern consisting of a single spike that actually appeared makes a high contribution to the number of relevant bins of all stimuli. On the other hand, patterns consisting of many spikes contribute much less to the number of relevant bins of each individual stimulus, and their total contribution to the number of relevant bins can be <1. Although this procedure can be considered as a variant of the 'Bayesian' estimate of the number of relevant bins, it was developed purely experimentally and its 'correctness' was judged only by the results on the simulated data.

(ii) *A decoding algorithm.* The decoding algorithm used here is due to Treves (Rolls et al., 1997). It is used to create a confusion matrix, in which each individual response is assigned to the most probable stimulus, where the probabilities of seeing a stimulus given a response are estimated using the other trials. The probability estimates for the number of spikes per time bin were parameterized as Gaussian (the Gaussian parametrization gave rise to better MI estimates in the IHPP simulations than the Poisson parameterization). Probabilities in different time bins were combined by assuming independence. The probability of observing a spike train given the stimulus was calculated using Bayes' equation. The method was applied to spike trains resampled at a number of temporal resolutions, as described in the Methods. MI estimates usually increased for higher temporal resolutions, and reached maximum with time bins of 2–4 ms. The maximum value, over all temporal resolutions, was used as the estimate of the MI between stimuli and responses.

This cross-validation procedure is a data reduction step (since the most probable stimulus is a function of the response). Hence the MI between rows and columns of the confusion matrix is, in principle, a lower bound on the MI between stimuli and responses (Bialek et al., 1991). However, with a good decoding algorithm, such a method can be highly efficient. Bias reduction schemes for this family of approaches are a special concern, since the confusion matrices do not necessarily fulfill the conditions for the standard bias correction schemes (Paninski, 2003). Indeed, in the parameter regime tested here, there was a tendency for over-estimation of the MI (see main text).

(iii) *2nd order series expansion.* We used the 2nd-order expansion of the MI in the window duration as described by Panzeri and Schultz (2001). The 1st order term is the MI due to the response rates alone. The 2nd order term can be separated into three contributions: the first depends on correlations created by the fact that different stimuli elicit different responses ('signal correlations'). It is always negative. The second term depends on the average correlation (over all stimuli) between responses to the same stimulus, and the third term depends explicitly on the difference in correlation structure between the responses to different stimuli. Both can be either positive or negative.

To apply this expansion, the rates and 2nd order temporal correlations must be estimated. When keeping spike times with very high time resolution, essentially all cross-products of instantaneous rates, required for computing the 2nd order temporal correlations, would be 0. To solve this difficulty, spike trains must be binned with finite bin width or convolved with a finite width window. However, binning or temporal smoothing constitutes a data reduction step, and will decrease the MI. We chose to widen the time bins, and the algorithm was applied to the resampled spike trains at all temporal resolutions used. The correlation coefficients were computed for each resampled train, and then the integration over time required by Panzeri's formulae was implemented as sums over time bins.

The Panzeri estimates turned out to be highly biased and highly variable. In fact, their values tended to be negative at the coarser temporal resolutions and to become positive (and too large) at the smallest temporal resolutions. The bias was estimated by averaging 10 estimates of the MI for datasets generated by randomly assigning responses to stimuli. For each temporal resolution, the bias was subtracted from the MI, and the highest estimate over all temporal resolutions was used.

(iv) *Binless MI estimation.* Victor (2002) suggested using a binless estimator of MI between stimuli and continuous data which seeks to minimize the loss of information due to binning. The method is based on the idea that the probability density of elements can be estimated using the distance of each element from its nearest neighbor. While this idea is commonly used for probability density estimation, Victor's method uses it to estimate the MI directly. This estimator approximates the ratio of probability densities between the conditional and unconditional response probabilities by the ratio of the distance between an observed response and its closest neighbor (i.e. the next most similar response) within the same stimulus set on the one hand, and across all stimuli on the other hand. The formula used by Victor is:

$$I = \frac{r}{N}\sum_{j=1}^{N}\log\frac{\lambda_j}{\lambda_j^*} - \sum_{k=1}^{s}\frac{N_k}{N}\log\frac{N_k - 1}{N - 1},$$

where r is the dimension of the measurement space (for example, $r = 1$ when the responses are quantified only in terms of mean response time), N is the total number of individual responses (e.g. $N = 960$ for the virtual space data and $N = 300$ for the cat data), N_k is the number of measurements of responses to stimulus k (for the ferret data, $N_k = 40$ and for the cat data $N_k = 20$), λ_j is the minimum distance between the jth response and its closest neighbor among all responses, and λ_j^* is the minimum distance between the jth response and its closest neighbor restricted to the responses to the same stimulus.

Although this formula is asymptotically unbiased, for the relatively limited sample sizes achievable in typical recording experiments it exhibits considerable bias, as will be shown below. In particular, we found that some responses make large negative contributions to the first term, when a given λ_j^* was very large, while the corresponding λ_j was very small. This happens when the jth data point was in a region of the response space that was not sampled well by the rest of the sample.

To estimate the bias we randomly assigned responses to stimuli and computed the MI of the resulting random set. The randomization process was repeated 10 times and the MI values averaged to produce the estimated bias. This procedure has no exact theoretical foundation for the binless estimates, but is expected to give reasonable results as long as the MI is far from the entropy of the stimulus set, as is the case here.

The formula developed by Victor can be directly applied to reduced response measures that produce continuous data, such as mean response time. However, for our data, estimates of mean response time information turned out to be more stable (and presumably less biased) when computed using histograms, despite the possible loss of information caused by the initial binning.

Victor's formula is not directly applicable to spike patterns, since these do not 'live' in an r-dimensional space. Victor suggested a two-step strategy. First, the entire set of spike trains is divided according to the spike count in each train. The MI can be expressed as (Victor, 2002, Eq. (14)):

$$I = I_{\text{count}} + \sum_{n=0}^{\infty} p(d(x) = n) I_{\text{timing}}(n),$$

where I_{count} is the MI due to the spike count distribution, $p(d(x) = n)$ is the probability of having n spikes in a train, and $I_{\text{timing}}(n)$ is the MI of the joint distribution of stimuli and responses, conditioned on having n spikes exactly in each response. In the second step, $I_{\text{timing}}(n)$ is evaluated by transforming the n-spike patterns into a D-dimensional pattern ($D <= n$) and using the binless estimate in the D-dimensional space. Note that the transformation is a data-reduction step, since D cannot be much larger than 3 or 4. Spike patterns with more than 3 or 4 spikes are therefore considered in a space of lower dimensionality. The utility of this method is therefore related to the contribution of higher-order patterns to the total MI. This contribution depends both on the probability of such patterns, which is relatively easy to estimate, and on the MI conditional on the presence of a large number of spikes, which is hard to estimate. It is therefore difficult to estimate the errors in this method.

The transformation we used is the embedding transformation suggested by Victor (2002). This transformation equalizes the marginal distributions across all D dimensions of the lower-dimensional space, and approximately orthogonalizes the coordinates. These features may not be crucial for the data presented here, and other embedding transformations may perform better. For example, embedding spike patterns of all sizes in a 1-dimensional space by using the mean response time would probably have the same effect as using the

2-dimensional reduced measure consisting of the pair of spike counts and mean response time, as done below.

Appendix III: Proof of the MI Decomposition Formula

Let $p(x, y)$ be the joint distribution of two random variables, X and Y. Suppose that X has a special value, denoted by n. The mutual information between X and Y can be written as

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{y} p(n,y)\log_2 \frac{p(n,y)}{p(n)p(y)} \\ &\quad + \sum_{x\neq n,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}, \\ &= A1 + A2 \end{aligned} \tag{A2}$$

where those terms that include the special value n are explicitly separated from the rest of the sum.

The term $A2$, that does not include the special value n, is now expressed in terms of the joint distribution of X and Y conditioned on X being different from n. The conditional joint distribution is

$$\bar{p}(x,y) = p(x,y)/p(x \neq n),$$

and its marginals are

$$\begin{aligned} \bar{p}(x) &= \sum_{y} \bar{p}(x,y) = p(x)/p(x \neq n), \\ \bar{p}(y) &= \sum_{x\neq n} \bar{p}(x,y) = p(x \neq n, y)/p(x \neq n). \end{aligned}$$

It follows that $p(x, y)/p(x) = \bar{p}(x, y)/\bar{p}(x)$, and therefore:

$$\begin{aligned} &\sum_{x\neq n,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x\neq n,y} p(x,y)\log_2 \frac{\bar{p}(x,y)}{\bar{p}(x)\bar{p}(y)} \cdot \frac{\bar{p}(y)}{p(y)} \\ &= \sum_{x\neq n,y} p(x,y)\log_2 \frac{\bar{p}(x,y)}{\bar{p}(x)\bar{p}(y)} \\ &\quad + \sum_{x\neq n,y} p(x,y)\log_2 \frac{\bar{p}(y)}{p(y)} \\ &= B1 + B2 \end{aligned} \tag{A3}$$

The term $B1$ in the last sum is in fact

$$\begin{aligned} &p(x \neq n) \sum_{x\neq n,y} \bar{p}(x,y)\log_2 \frac{\bar{p}(x,y)}{\bar{p}(x)\bar{p}(y)} \\ &= p(x \neq n) I(X(x \neq n); Y), \end{aligned}$$

and is therefore the residual information in the joint distribution of X and Y, conditioned on X being different from the special value n.

The term $B2$ can be rewritten as:

$$\begin{aligned} &\sum_{x\neq n,y} p(x,y)\log_2 \frac{\bar{p}(y)}{p(y)} \\ &= \sum_{y} p(x \neq n, y)\log_2 \frac{p(x \neq n, y)}{p(y)p(x \neq n)}. \end{aligned}$$

It can be combined now with the term $A1$ from Eq. (A2):

$$\begin{aligned} A1 + B2 &= \sum_{y} p(n,y)\log_2 \frac{p(n,y)}{p(y)p(n)} \\ &\quad + \sum_{y} p(x \neq n, y)\log_2 \frac{p(x \neq n, y)}{p(y)p(x \neq n)}. \end{aligned}$$

This is in fact the MI between a random variable that gets two values, one (e.g. 0) when $X = n$ and the other (e.g. 1) otherwise, and Y. This is the variable z as defined in the Methods section.

Acknowledgments

This work was supported by a Biotechnology and Biological Sciences Research Council grant to J. W. H. Schnupp, a Wellcome Trust studentship to T. D. Mrsic-Flogel, a Wellcome Trust Senior Research Fellowship to A. J. King, by a grant from the Israeli ministry of science to G. Chechik, and by a Volkswagen grant and a GIF grant to I. Nelken.

References

Bar-Yosef O, Rotman Y, Nelken I (2002) Responses of neurons in cat primary auditory cortex to bird chirps: Effects of temporal and spectral context. J. Neurosci. 22: 8619–8632.

Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D (1991) Reading a neural code. Science 252: 1854–1857.

Brugge JF, Reale RA, Hind JE (1996) The structure of spatial receptive fields of neurons in primary auditory cortex of the cat. J. Neurosci. 16: 4420–4437.

Brugge JF, Reale RA, Jenison RL, Schnupp J (2001) Auditory cortical spatial receptive fields. Audiol Neurootol. 6: 173–177.

Cover T, Thomas J (1991) Elements of Information Theory. Wiley and Sons, NY.

DeGroot MH, Schervish MJ (2001) Probability and Statistics. Addison-Wesley, Boston, MA.

Duda RO, Hart PE, Stork DG (2000) Pattern Classification, 2nd edition. John Wiley and Sons.

Furukawa S, Middlebrooks JC (2002) Cortical representation of auditory space: Information-bearing features of spike patterns. J. Neurophysiol. 87: 1749–1762.

Jenison RL, Reale RA (2003) Likelihood approaches to sensory coding in auditory cortex. Network 14: 83–102.

Johansson RS, Birznieks I (2004) First spikes in ensembles of human tactile afferents code complex spatial fingertip events. Nat Neurosci. 7: 170–177.

Machens CK, Schutze H, Franz A, Kolesnikova O, Stemmler MB, Ronacher B, Herz AV (2003) Single auditory neurons rapidly discriminate conspecific communication signals. Nat Neurosci. 6: 341–342.

Middlebrooks JC, Clock AE, Xu L, Green DM (1994) A panoramic code for sound location by cortical neurons. Science 264: 842–844.

Middlebrooks JC, Xu L, Eddins AC, Green DM (1998) Codes for sound-source location in nontonotopic auditory cortex. J. Neurophysiol. 80: 863–881.

Mrsic-Flogel TD, King AJ, Schnupp JWH (2005) Encoding of virtual acoustic space stimuli by neurons in ferret primary auditory cortex. J. Neurophys. 93: 3489–3503.

Paninski L (2003) Estimation of entropy and mutual information. Neural. Comput. 15: 1191–1253.

Panzeri S, Treves A (1996) Analytical estimates of limited sampling biases in different information measures. Network 7: 87–101.

Panzeri S, Schultz SR (2001) A unified approach to the study of temporal, correlational, and rate coding. Neural Comput. 13: 1311–1349.

Panzeri S, Schultz SR, Treves A, Rolls ET (1999a) Correlations and the encoding of information in the nervous system. Proc. R. Soc. Lond B. Biol. Sci. 266: 1001–1012.

Panzeri S, Treves A, Schultz S, Rolls ET (1999b) On decoding the responses of a population of neurons from short time windows. Neural Comput. 11: 1553–1577.

Panzeri S, Petroni F, Petersen RS, Diamond ME (2003) Decoding neuronal population activity in rat somatosensory cortex: Role of columnar organization. Cereb. Cortex 13: 45–52.

Panzeri S, Petersen RS, Schultz SR, Lebedev M, Diamond ME (2001) The role of spike timing in the coding of stimulus location in rat somatosensory cortex. Neuron 29: 769–777.

Petersen RS, Panzeri S, Diamond ME (2001) Population coding of stimulus location in rat somatosensory cortex. Neuron 32: 503–514.

Petersen RS, Panzeri S, Diamond ME (2002) The role of individual spikes and spike patterns in population coding of stimulus location in rat somatosensory cortex. Biosystems 67: 187–193.

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.

Reale RA, Jenison RL, Brugge JF (2003) Directional sensitivity of neurons in the primary auditory (AI) cortex: Effects of sound-source intensity level. J. Neurophysiol. 89: 1024–1038.

Reich DS, Mechler F, Victor JD (2001) Formal and attribute-specific information in primary visual cortex. J. Neurophysiol. 85: 305–318.

Rieke F, Warland D, deRuyter van Steveninck R, Bialek W (1997) Spikes. MIT Press, Cambridge, MA.

Rolls ET, Treves A, Tovee MJ (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. Exp. Brain Res. 114: 149–162.

Rotman Y, Bar-Yosef O, Nelken I (2001) Relating cluster and population responses to natural sounds and tonal stimuli in cat primary auditory cortex. Hear Res. 152: 110–127.

Schervish MJ (1995) Theory of Statistics. Springer, New York.

Schnupp JWH, Mrsic-Flogel TD, King AJ (2001) Linear processing of spatial cues in primary auditory cortex. Nature 414: 200–204.

Slonim N, Tishby N (2000) Agglomerative Information Bottleneck. In: (SA Solla, TK Leen, KR Muller, eds), Advances in Neural Information Processing Systems 12. MIT Press, Cambridge, MA.

Sokal RR, Rohlf FJ (1981) Biometry, 2nd edition. W.H. Freeman, New York.

Treves A (2001) Information Coding in Higher Sensory and Memory Areas. In: F Moss, S Gielen, eds., Handbook of Biological Physics, Vol. 4: Neuro-Informatics and Neural Modelling. Elsevier, Amsterdam, pp. 825–852.

Treves A, Panzeri S (1995) The upward bias in measures of information derived from limited data samples. Neural Computation 7: 399–407.

Van Rullen R, Thorpe SJ (2001) Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. Neural Comput. 13: 1255–1283.

Van Rullen R, Thorpe SJ (2002) Surfing a spike wave down the ventral stream. Vision Res. 42: 2593–2615.

Van Rullen R, Gautrais J, Delorme A, Thorpe S (1998) Face processing using one spike per neurone. Biosystems 48: 229–239.

Victor JD (2000) Asymptotic bias in information estimates and the exponential (Bell) polynomials. Neural Comput. 12: 2797–2804.

Victor JD (2002) Binless strategies for estimation of information from neural data. Phys. Rev. E 66: 51903.

Victor JD, Purpura KP (1996) Nature and precision of temporal coding in visual cortex: A metric-space analysis. J. Neurophys. 76: 1310–1326.