Available online at www.sciencedirect.com
ScienceDirect
Hearing Research 229 (2007) 94–105
Hearing Research
www.elsevier.com/locate/heares

Research paper

Information theory in auditory research

Israel Nelken [a,*], Gal Chechik [b]

[a] Department of Neurobiology and The Interdisciplinary Center for Neural Computation, Silberman Institute of Life Sciences, Safra Campus, Givat Ram, Hebrew University, Jerusalem 91904, Israel
[b] Computer Science Department, Stanford University, Stanford CA 94305, USA

Received 10 September 2006; received in revised form 22 November 2006; accepted 3 January 2007
Available online 16 January 2007

Abstract

Mutual information (MI) is in increasing use as a way of quantifying neural responses. However, it is still considered with some doubts by many researchers, because it is not always clear what MI really measures, and because MI is hard to calculate in practice. This paper aims to clarify these issues. First, it provides an interpretation of mutual information as variability decomposition, similar to standard variance decomposition routinely used in statistical evaluations of neural data, except that the measure of variability is entropy rather than variance. Second, it discusses those aspects of the MI that makes its calculation difficult. The goal of this paper is to clarify when and how information theory can be used informatively and reliably in auditory neuroscience.

Keywords: Auditory system; Information theory; Entropy; Mutual information; Variance decomposition; Neural code

1. Introduction

In recent years, information-theoretic measures are increasingly used in neuroscience in general, and in auditory research in particular, as tools for studying and quantifying neural activity. Measures such as entropy and mutual information (MI) can be used to gain deep insight into neural coding, but can also be badly abused. This paper is an attempt to present those theoretical and practical issues that we found particularly pertinent when using information-theoretic measures in analyzing neural data.

The experimental context for this paper is that of measuring a stimulus–response relationship. In a typical experiment, a relatively small number of stimuli ($\sim$<100) are presented repeatedly, typically 1–100 repeats for each stimulus. The main experimental question is whether the neuronal activity was different in response to the different stimuli. If so, it is concluded that the signal whose activity is monitored (single-neuron responses, evoked potentials, optical signals, and so on) was selective to the parameter manipulated in the experiment.

The MI is a measure of the strength of association between two random variables. The MI, $I(S;R)$, between the stimuli S and the neural responses R is defined in terms of their joint distribution $p(S,R)$. When this distribution is known exactly, the MI can be calculated as

$$I(S;R) = \sum_{s\in S, r\in R} p(s,r)\log_2\left(\frac{p(s,r)}{p(s)p(r)}\right)$$

where $p(s) = \sum_{r\in R} p(s,r)$ and $p(r) = \sum_{s\in S} p(s,r)$ are the marginal distributions over the stimuli and responses, respectively.

The easy way to use the MI is to test for significant association between the two variables. Here the null hypothesis is that the two variables are independent. The distribution of the MI under the null hypothesis is (with appropriate scaling) that of a χ^2 variable, leading to a significance test for the presence of association (e.g. Sokal and Rohlf, 1981; where it is called the G-statistic). Using the MI in this way, only its size relative to the critical value of the test is of importance.

* Corresponding author. Tel.: +972 2 6584229; fax: +972 2 6586077.
E-mail address: israel@cc.huji.ac.il (I. Nelken).

0378-5955/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.heares.2007.01.012

A more complicated way of using the MI is to try to estimate its actual value, in which case it is possible to make substantially deeper inferences regarding the relationships between the two variables. This estimation is substantially more difficult than performing the significance test. The reasons to undertake this hard estimation problem, and the associated difficulties, are the main subject of this paper.

2. Why mutual information?

2.1. The Mutual Information as a measure of stimulus effect

Neuronal responses are high-dimensional: to fully characterize in detail any single spiking response to a stimulus presentation, it is necessary to specify many values, such as the number of spikes that occurred during the relevant response window and their precise times. Similarly, membrane potential fluctuates at >1000 Hz, and therefore more than 200 measurements are required to fully specify a 100 ms response. We usually believe that most of the details in such representations are unimportant, and instead of specifying all of these values, typically a single value is used to summarize single responses – for example, the total spike count during the response window, or first spike latency, or other such simple measures, that will be called later 'reduced measures' of the actual response.

Having reduced the representation of the responses to a single value, it is now possible to test whether the stimuli had an effect on the responses. Usually, the effect that is tested is a dependence of the firing rate of the neuron on stimulus parameters. For example, to demonstrate frequency selectivity, we will look for changes in firing rates of a neuron as a function of tone frequency.

To understand what information-theoretic measures tell us about neuronal responses, let us consider the standard methods for performing such tests in detail. A test for a significant difference between means is really about comparing variances (Fig. 1): the variation between response means has to be large enough with respect to variation between responses to repeated presentations of the same stimulus.

Initially, all the responses to all stimuli are pooled together, and the overall variability is estimated by the variance of this set of values around its grand mean. Fig. 1 shows the analysis of artificial data that represents 20 repeats of each of two stimuli (these are actually samples of two Poisson distributions with expected values of 5 and 10). In Fig. 1a, the overall distribution of all responses (both of stimulus 1 and of stimulus 2) is presented. The total variance is 10.9 (there are no units, since these are spike counts), corresponding to a standard deviation of about 3 spikes.

Part of the overall variation occurs because responses to repeated presentation of the same stimulus are noisy – this is called *within-stimulus* variability. Another part of this overall variation is due to the fact that different stimuli cause different responses. A stimulus effect is significant if

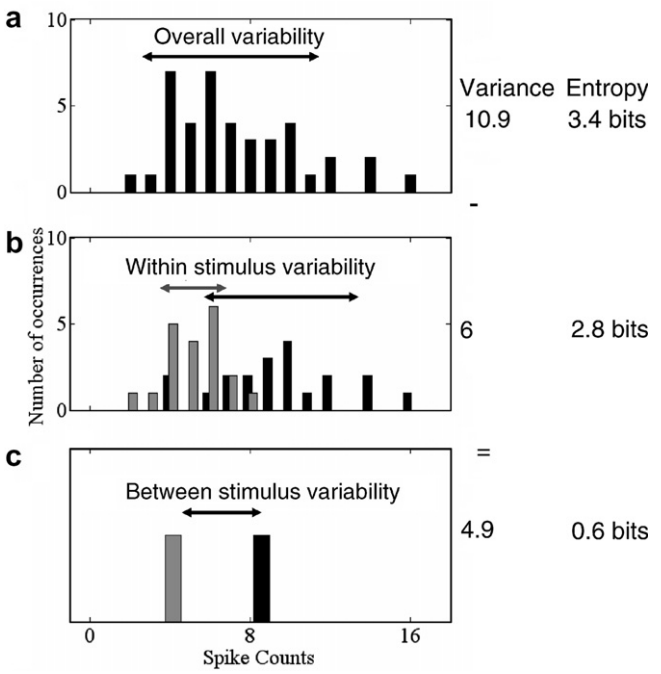

Fig. 1. Variability analysis using variances and entropies. (a) Overall distribution of 40 measurements (2 stimuli, 20 repeats of each stimulus). (b) Distribution of the responses to the two stimuli. (c) Samples means.

the second variability source, *between-stimulus* variation, is large enough relative to the first variability source, the within-stimulus variability. Conceptually, the next step is to compute within-stimulus variability. To do that, the variance of all responses to each stimulus, around their own mean, is computed (Fig. 1b). The two histograms represent the responses to stimulus 1 (black) and stimulus 2 (gray), with variances of 10.1 and 2 (standard deviations of about 3 and 1.5 spikes). This set of variances is then averaged across stimuli, and used as an estimate of the within-stimulus variability – for the data in Fig. 1, the average within-stimulus variance is about 6.

It can be shown mathematically that within-stimulus variability will always be smaller than the overall variance, and the difference between them is the variability between the means of the responses to the different stimuli (Fig. 1c). Thus, the goal of dividing variance into two sources, the within-stimulus variance and the across-stimulus variance, is achieved.

This decomposition has good statistical properties, in the sense that the two variability sources are uncorrelated. Statistical theory can now be used to determine when the ratio between the two variability sources should be considered as larger than expected under the assumption of no stimulus effect (Sokal and Rohlf, 1981), leading to specific statistical tests (e.g. the *F*-test of the 1-way ANOVA) in Fig. 1 the *F*-test (or the equivalent *t*-test for equality of means, which is essentially the same thing here) comes out highly significant.

The recipe given above is extremely powerful, and therefore unsurprisingly is extensively used. However, it has

some serious limitations. For example, could it be that using more detailed descriptions of the responses, a stronger stimulus effect could be found? Moreover, in some cases the use of spike counts is inappropriate, and other reduced measures such as first spike latency should be used. This is the case for example, if a neuron usually spikes always once, but at a different point in time depending on the stimulus. However, the issue here is general: How do we know the reduced measure we use is the best, or even that it makes sense?

Even more importantly, this procedure can only be followed when means and variances can be computed. This is true when responses are summarized by numerical-valued variables such as spike counts or simple measures of spike timing, but is problematic in other situations. Nominal or ordinal variables cannot be analyzed by this procedure. More importantly, neuronal responses in their full complexity cannot be analyzed in this framework. For example, the mean and the variance of a set of precise spike patterns are difficult to define in a natural way: such definition requires many assumptions about those elements of the spike patterns that are important for coding, the underlying noise structure and the relevant temporal resolution. Finally, the mean and variance capture only some aspects of the distribution, and may ignore other aspects of the responses that could also encode properties of the stimulus.

We would like to keep the general framework of variability decomposition, without using variances. A solution is provided by information theory, supplying a different measure of variability – the *entropy* of the probability distribution (Cover and Thomas, 1991). The entropy is defined as

$$H(p) = \sum_i p(i)\log_2\left(\frac{1}{p(i)}\right),$$

where $p(i)$ are the probabilities of all the different values that the random variable can have (here we assume a discrete random variable in order to avoid the technicalities associated with calculation of entropies for continuous distributions).

The entropy does not assume anything about the relationships between different possible values that can be achieved, and therefore can always be computed, even for nominal or ordinal variables, when means and variances do not make sense. The entropy has many of the properties of variance – it is non-negative, and it is equal to 0 only when the random variable has a single value with probability 1. In other respects, entropy and variance are different. For example, scaling a variable would change its variance, but not its entropy. Thus, a variable having two values with probability 0.5 each would have entropy of 1 bit whether the two values are -1 and 1 or -10^{17} and 10^{17}. Therefore, the entropy codes different aspects of variability than the variance.

Using entropy, it is possible to perform variability decomposition in the same way as with variance. First, compute the overall response entropy. For the data in Fig. 1a, the overall response entropy is 3.4 bits. Next, compute the entropy of the responses for each stimulus separately and average across stimuli. In Fig. 1b, the entropy is 3.2 and 2.5 bits for the distribution of counts of stimuli 1 and 2, respectively, and their average is 2.8 bits. This number is called the *conditional entropy* – in this case the entropy of the responses conditioned on the stimuli. It is a measure of the variability of different responses to the same stimulus. The difference between the overall entropy and the conditional entropy should reflect the effect of the stimulus – in Fig. 2 it is 0.6 bits. In fact, this difference is precisely the MI. Thus, the MI is a measure of 'stimulus effect' – that part of total response variability (measured by entropy) that is due to difference between responses to the different stimuli. Again, for a test of association, the MI in this case is highly significant.

2.2. Useful properties of the MI

If the MI is not more than another way to do variance decomposition, why use it at all? The MI has theoretical properties that make it ideal for addressing some general questions in neuroscience. I will highlight three such properties here.

2.2.1. Symmetry

The defining formula of the MI is symmetric in the stimuli and the responses. On the other hand, the recipe above for calculating the MI as part of a variability decomposition is asymmetric, because stimuli and responses do not play the same role: the MI was computed as the difference

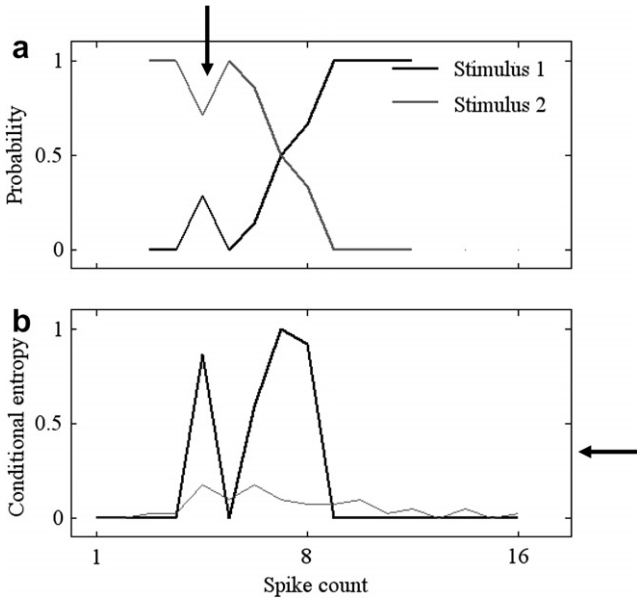

Fig. 2. Computing the MI by conditioning on responses. (a) The conditional distributions of stimuli given the responses. Black line – probability of stimulus 1; gray line – probability of stimulus 2. (b) The resulting conditional stimulus entropies, one value for each possible response. Gray – the weights used for computing the average.

between overall response entropy and the average response entropy computed for the single stimuli. However, a similar calculation can be done with the role of stimuli and responses reversed.

To do this, the entropy of the stimuli is computed first. Since the distribution of stimuli is part of the experimental design, stimulus entropy is under complete experimental control. For the data in Fig. 1, with two equiprobable stimuli, the stimulus entropy is 1 bit. Next, individual stimulus repetitions are assigned to classes according to the responses they evoked (Fig. 2). Thus, if the response is quantified by spike counts, stimuli are classified by the number of spikes they evoked. For example, part of this calculation consists of selecting all stimuli that evoked 4 spikes. Of the 40 stimulus presentations (20 of stimulus 1 and 20 of stimulus 2), 7 stimulus presentations evoked 4 spikes. Of these, 2 were stimulus 1 and 5 were stimulus 2. As a result, conditional on having evoked 4 spikes, the probability of stimulus 1 is estimated as 2/7 and that of stimulus 2 is estimated as 5/7. This is nothing but a direct application of Bayes rule. Fig. 2a shows these distributions for all observed spike counts (the case of 4 spikes is indicated with a vertical arrow).

Next the entropy of these probability distributions is computed and averaged across responses – this is the conditional stimulus entropy, conditioned on observing the responses. These are displayed in Fig. 2b, and their average, the conditional stimulus entropy, is about 0.4 bits (marked with an horizontal arrow – the weights used for the different responses are marked in gray). Finally, the difference between the overall stimulus entropy and this averaged within-response stimulus entropy can be computed, giving a measure of 'response effect', which is again 0.6 bits.

It turns out that the two ways of calculating an effect, starting with responses or starting with stimuli, always result in the same number, the MI as defined above. This is reflected in the symmetric way in which the two variables participate in the definition of the MI. Thus, the MI can be interpreted in two ways. It is stimulus effect on the responses – the part of response variability that is due to variation in stimuli. But it is also the response effect on the stimuli – the part of the variability in the stimuli that is accounted for by observing responses. Whereas the first view is that of encoding – how responses encode stimuli – the second view is that of decoding – to what extent the stimulus can be determined after observing a response. The identity of the resulting numbers creates a deep link between encoding and decoding.

2.2.2. Scale

Since conditioned entropies are positive but smaller than overall entropies, the MI is always non-negative but will always be smaller than both the entropy of the responses and the entropy of the stimuli (in fact, the entropy of the stimulus $H(S)$ is mathematically equivalent to $I(S;S)$, and similarly for $H(R)$; furthermore, $I(S;S) >= I(S;R)$, which is a consequence of the information processing inequality to be discussed below). We can therefore know when the MI is small and when it is large by comparing it with the smaller between stimulus and response entropies.

When the MI is zero, stimuli and responses are independent – there is no effect whatsoever of the stimuli on the responses. This is a far stronger statement than the statement that there is no stimulus effect on response *means* – the MI is sensitive to all possible departures from independence (Fig. 3 illustrates some possibilities: unequal means, equal means and unequal variance, and even equal means and equal variances but small differences in the detailed distributions). Therefore when $MI = 0$, any test, on any measure of the responses, will not be able to uncover a significant stimulus effect. Symmetrically, any decoder of the responses will perform at chance level.

Fig. 3 may also serve to calibrate the expectations for the size of the MI. Because of the absolute scale, in the case of two stimuli the MI cannot be larger than 1 bit. In the case illustrated in Fig. 3a, the MI is only 0.44 bits, which

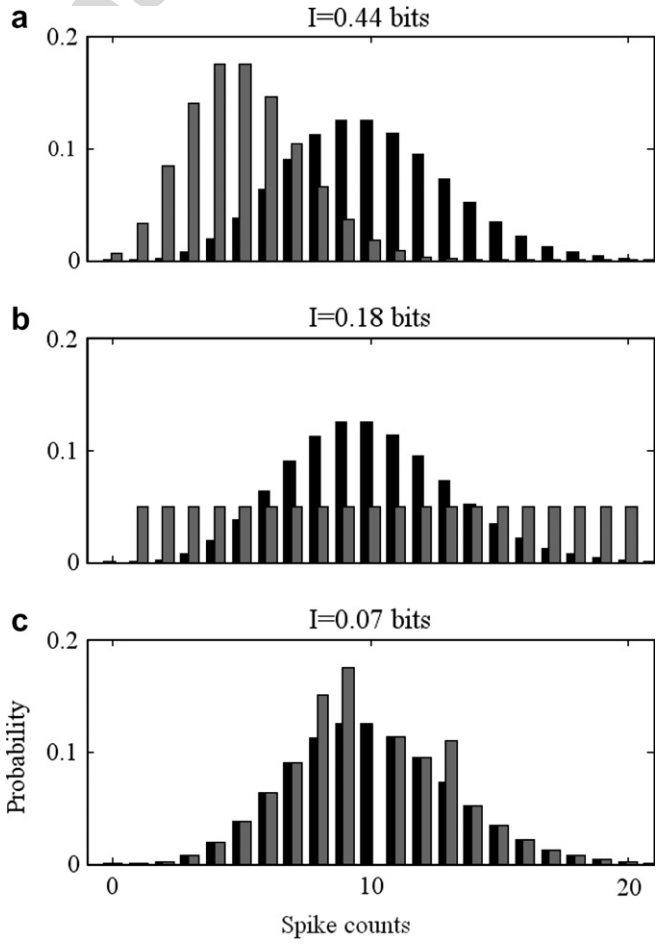

Fig. 3. The MI is sensitive to general departures from independence. (a) Two Poisson distributions, different means and different variances. (b) A Poisson distribution and a uniform discrete distribution with the same mean. Whereas an ANOVA test would most probably come not significant, the MI is non-zero (although not very large). (c) Two distributions with the same mean and the same variance. The MI is now small, but still non-zero.

is somewhat less than a half of the maximum possible value, although the two distributions compared there would be considered as highly discriminable experimentally.

In fact, if the MI is equal to stimulus entropy (1 bit in Fig. 3), the stimuli can be perfectly recovered from the responses. This is so because the response-conditioned entropy in this case is 0, which can happen only if each response perfectly specifies a single stimulus (although a given stimulus can in principle still evoke different responses on different presentations). In terms of the calculations illustrated in Fig. 2, the conditional probabilities of the stimuli given the responses (Fig. 2a) should all be either 0 or 1, never in between.

These absolute bounds have other implications as well. For example, in order to decode the responses well, we would like the MI to be close to *stimulus* entropy. However, the MI cannot be larger than *response* entropy. Thus, if response entropy is smaller than stimulus entropy, perfect decoding is impossible. Response entropy depends to some extent on experimenter choices. When responses are summarized with more details, response entropy will typically increase. Thus, if responses are summarized by whether a neuron responded or not, the maximum response entropy is 1 bit. If the response is summarized by the number of evoked spikes, response entropy can be larger, depending on the possible spike counts and their distribution. In this sense, more detailed descriptions of the responses are 'good' – some of the increased overall entropy might be used to encode significant variability between stimuli. Of course, increased details may also backfire, as we will discuss below.

2.2.3. The information processing inequality

Roughly speaking, the information processing inequality says that any processing done on either stimuli or responses will at best keep the same level of MI, or may lose MI. In other words, it is impossible to gain information by processing data, only to lose information.

This mathematical result needs to be interpreted with care. A sound has many properties, some of which may be important and some not. For example, when recording responses of an auditory nerve fiber to a low-frequency pure tone, the absolute times of the spikes relative to stimulus onset carry information about its phase – starting the stimulus at different phases will result in reproducible changes in absolute spike timing (simulated in Fig. 4a and b). This information is highly important for computing interaural phase differences, but may be irrelevant in other cases (e.g. when doing a frequency discrimination task), and can be safely ignored in order to better specify other aspects of the response (e.g. the autocorrelation of the spike train emphasizes the periodical structure, losing phase information, see Fig. 4c). Thus, overall information loss is often accepted when readout of relevant aspects of the sound becomes easier. Because of the practical aspects of MI estimation discussed below, such data processing

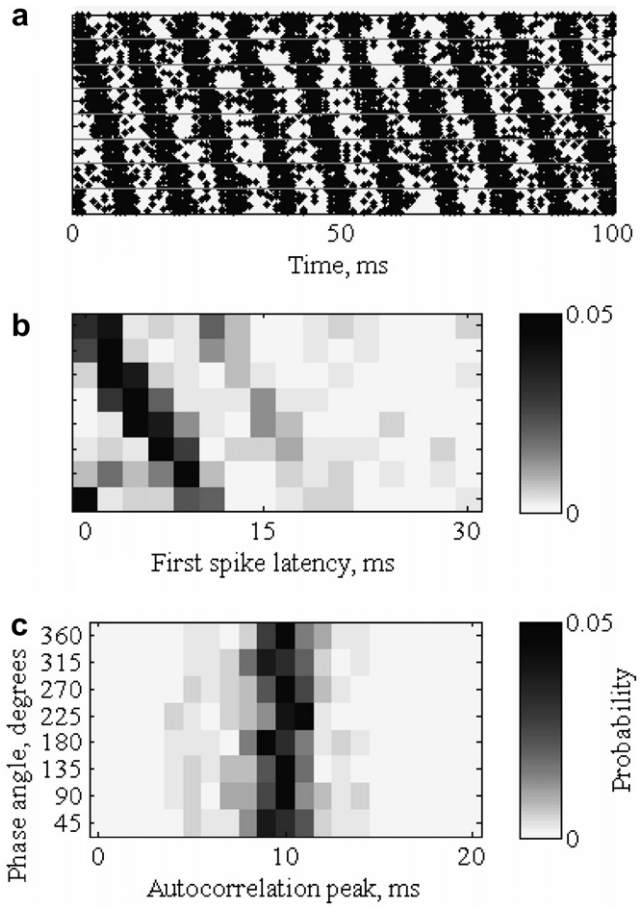

Fig. 4. Processing for enhancing specific aspects of stimulus representation. (a) Raster plots for simulated ANF in response to a 100 Hz tone. The phase of the tone was shifted, and each value of the phase was used for 50 responses. (b) Joint distribution of stimuli and first spike latency. Note that the joint distribution has a lot of structure, leading to a substantial MI (about 0.7 bits). (c) Location of the first peak of the single-trial autocorrelation function. The phase sensitivity has been abolished by this analysis, but the periodicity is much more strongly represented.

may actually increase the amount of information that can be extracted about the frequency of the stimulus under typical experimental conditions.

The importance of the information processing inequality, however, goes far beyond these general comments. It is a basic theoretical tool for discussing decoding of neural activity.

One of the most important techniques used to quantify stimulus–response relationships has been to build a decoder – an algorithm whose input is an experimentally-measured response and whose output is a guess at the stimulus that evoked the response. In order to demonstrate significant stimulus–response associations, it is enough to build a decoder functioning significantly above chance level. A decoder is considered as a proof of principle that a (half-mythical) 'next layer' could extract relevant information about the world from the neuronal responses.

Having a good decoder is however just a first step. We want now to compare neurons by the performance of

decoders trained on their responses; or to compare two different sets of stimuli to demonstrate that a neuron is more sensitive to one or to the other; or to build decoders based on different features of the responses, in order to find out whether these response features carry information about the stimuli. However, a decoder is usually coming from a specific class of algorithms (multilayer perceptrons, or support vector machines, or radial basis neural networks, or other such families) and is trained on the data using a specific set of parameters. One could imagine that using a different class of algorithms, or even different parameter settings for the same algorithm, could result in a better decoder, in which case any comparison between decoders working on different data is invalid. In other words, any specific decoder only bounds from below the performance of the best decoder for the task we study.

Is it possible to improve a decoder beyond any given bound by using better and better algorithms? The answer is no – the information processing inequality puts an absolute bound on decoder performance. Decoders are quantified by their transmitted information: this is the MI between the true stimuli and the decoded ones. However, the decoded stimuli are a function of the responses – this is precisely the meaning of a decoder: a machine that gets responses and guesses what were the stimuli. Therefore, the transmitted information of any decoder, being the mutual information between stimuli and a function of the responses, cannot be larger than the MI between stimuli and (the unprocessed) responses. In this respect, training a decoder is a way of computing a lower bound (the transmitted information) on the MI between stimuli and responses. Alternatively, a valid estimate of the MI between stimuli and responses bounds the performance of any decoder. Therefore, when looking for an optimal decoder, it makes sense to spend the effort on computing a valid estimate of the MI between stimuli and responses, which is an absolute bound with respect to which the transmitted information of any decoder should be compared.

A decoder that reaches the bound set by the information processing inequality is a function f of the data for which $I(S; f(R)) = I(S; R)$. Such a decoder may or may not exist, but functions of the responses that do not lose information are interesting in their own right. This is actually a rather large family. For example, any function that associates a unique value in its range to each response value has this property, since in that case the relationship between R and $f(R)$ can be inverted. This is what we do when we code 0–1 spike patterns as the numbers they denote in binary notation. However, such functions are not necessarily useful, since they do not simplify the responses in any way; they just replace complexity with another complexity. On the other hand, a perfect decoder is a very useful member of this family.

Note that in this respect, the actual values of the function $f(R)$ are unimportant. The important action of f is to identify possible responses: those responses that give rise to the same value of f. By identifying these responses, using f is tantamount to accepting that the differences between them are not important (think again of using spike counts instead of exact spike patterns). Thus, in information-theoretic terms, the main effect of f is to reduce the variability (as measured by entropy) of the response space. Useful information-preserving functions reduce the complexity of the response space, but keep those aspects of response variability that are important for coding the stimuli.

Functions that keep information are therefore candidates for being the neural code, in the sense that the 'next layer' in the brain does not need to know all the details of the response – it is enough to know the value of one of these functions in order to have access to the full information about stimuli. Used in this way, the information processing inequality is a rigorous tool to look for candidates for the neural code.

In the same way, the information processing inequality can be used to learn about features of stimuli that are important for shaping the responses of complex neurons. In this case, the stimuli have a complex description, and we are looking for a reduced measure of the stimuli that would keep the mutual information with the responses. Such a function encapsulates the relevant information about the stimulus that is necessary in order to fully specify the response. A similar argument was used by Sharpee et al. (2004) to find the so called maximally informative dimensions – linear filters in stimulus space that keep maximal information about the responses.

2.3. Other interpretations of MI

(i) The MI is the reduction in uncertainty about the stimulus after a single response is observed. This is the standard information-theoretic interpretation. In the context studied here, without observing the response, the guess which stimulus was presented must be solely based on the experimental design, and is quantified by stimulus entropy. For example, with 16 equiprobable stimuli, the guess will select each stimulus with probability 1/16 and the entropy is 4 bits. If the mutual information between a neuron and the stimuli is e.g. 0.5 bit, observing the responses of the neuron reduces this entropy to 3.5 bits on average. The averaging here is important – one can easily construct cases in which observation of a specific response increases uncertainty about the stimulus. For example, imagine an experiment with two stimuli, one of which has a probability of 0.1 and the other 0.9. Stimulus entropy is in this case is about 0.47 bits (instead of 1 bit for equiprobable stimuli). Now suppose that the joint distribution of stimuli and responses is such that there is a low-probability response that is 9 times more probable when the low probability stimulus is presented than when the high probability stimulus is presented. Under these circumstances, the conditional probability of the two stimuli after observing that specific response is 0.5 so that the conditional entropy is

1 bit, higher than the original stimulus entropy. However, in that case other response values will have lower stimulus entropy since we know that on average the conditional stimulus entropy will be smaller than the a-priori stimulus entropy, 0.47 bits.

With this interpretation, the MI can be used to estimate the number of neurons that are required to decode stimuli perfectly on a trial-by-trial basis. The estimate is the stimulus entropy divided by the single-neuron MI. However, this estimate is not guaranteed to bound the number of neurons either from above or from below, since the MI can add supra-linearly or sub-linearly among neurons. How information of ensembles relates to single-neuron MI is a complex question, outside the scope of this review (see Chechik et al., 2006; Deneve et al., 2001; Nirenberg and Latham, 2003; Schneidman et al., 2003).

(ii) The MI is the log (to the base of 2) of the number of different classes to which the stimuli can be subdivided after observing a response. This interpretation is tightly linked to the previous one, and is a concrete interpretation of the reduction in uncertainty. However, single-neuron MI is often smaller than 1 bit, making the number of different classes smaller than 2. Thus, this interpretation seems less useful for single-neuron calculations.

2.4. Practice of mutual information estimation

If the MI is so useful, why is not it used more often? One answer to this question is that the MI has entered the field late, and that it is gaining respect as a tool for analyzing neural data. However, there is another good reason for the limited use of information-theoretic measures: calculating the actual value of the MI between two experimentally measured variables is theoretically difficult, and the difficulty translates into practical hurdles.

In principle, the MI is one value that summarizes some properties of a probability distribution. As such, it seems that it would not be more difficult to estimate than e.g. a mean and a variance (Nemenman et al., 2004). This is however incorrect – the estimation of the MI is a hard problem. Theoretical aspects of MI estimation have been most rigorously treated by Paninski (2003), who highlighted the severe problems encountered when trying to estimate MI from data.

Estimating MI from empirical data commonly involves two steps: first, estimating the joint distribution of stimuli and responses, and then calculating the MI based on this estimated distribution. The first step in such calculations requires estimating the distribution of neural responses for each stimulus. For example, when interested in information in spike counts, one calculates the distribution of number of spikes in the responses, as measured across repeated presentation of each one of the stimuli separately. Repeating this calculation for each stimulus yields the joint distribution of stimuli and responses. Fig. 5 is an illustration of this procedure. In this example, two stimuli are presented with equal probability. In fact, these are the spike trains that were used to generate the data of Fig. 1. They are samples of two homogeneous Poisson processes, so that the spike counts are sufficient statistics and follow a Poisson distribution, as stated above. The histograms of the spike counts are given in Fig. 1b, and they are normalized to give a joint distribution in Fig. 5b. For example, $\tilde{p}(0,1)$ is the probability of a response to have no spikes and be a response to stimulus 1 (Fig. 5b). Other statistics of spike patterns can be used instead of spike counts. For example, spike trains can be viewed as binary "words" of some fixed length. In Fig. 5c, the spike trains of Fig. 5a were divided into 5 periods of 20 ms each, and recoded as 5-bit binary words, where each bit is 1 if at least one spike occurred during the appropriate period. The distribution of the binary words can be estimated by counting the number of appearances of each word across repeated presentations of each stimulus. In our case, the spike rate evoked by stimulus 1 was such that in a large number of trials, at least one spike appeared in each of the periods, making

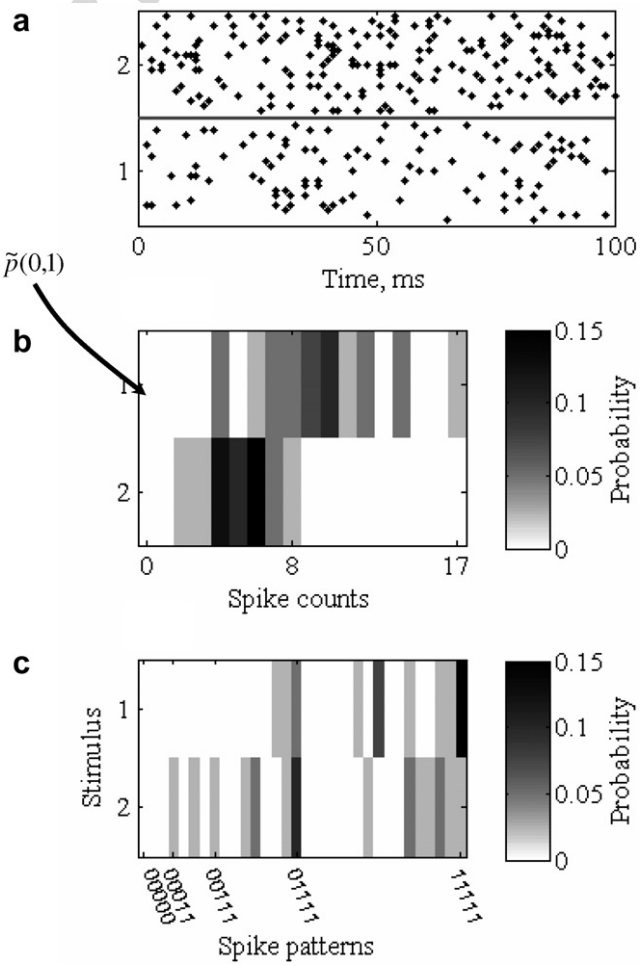

Fig. 5. Calculation of the MI. (a) rasters of the responses to two stimuli. (b) Representation in gray scale, after normalization, of (B). (c) Joint distribution of spike patterns and stimuli. Here, a pattern is defined as a summary of 5 periods of 20 ms each, coded as 1 if at least one spike occurred and 0 otherwise.

the full pattern '11111' by far the most common one. On the other hand, at least one spike occurred in each presentation, making the estimated probability of the spike pattern '00000' zero.

The second step is to calculate MI from the joint distribution. When the number of samples is very large relative to the number of bins in the joint distribution matrix, the observed empirical joint distribution provides a good estimate of the true underlying distribution, and the MI can be calculated by plugging in the empirical distribution. Unfortunately, with common experimental settings the number of samples is often not sufficient, and this naive MI estimator is positively biased. This means that it tends to produce overestimates of the MI relative to the MI of the true distribution,

$$I(\tilde{p}(S,R)) > I(p(S,R))$$

(but not always: the theoretical bias may be negative, see the discussion in Paninski, 2003). In addition, the variability of the estimator due to finite sampling may be also considerable. It has been shown that a first order approximation of the expected bias is

$$\frac{\#bins}{2N\log(2)},$$

where *#bins* is the number of independent 'essentially non-zero' bins in the joint distribution (those that potentially might be non-zero; if actually zero, this is due to finite sampling) and N is the number of samples (Panzeri and Treves, 1996; Treves and Panzeri, 1995). Subtracting this estimate of the expected bias from the empirical MI estimate often reduces substantially the bias.

As an example, the spike counts for the data in Fig. 1a are actually a sample from the Poisson distributions of Fig. 3a, whereas the naïve MI estimate is about 0.60 bits, had we had the full distributions (by e.g. repeating the stimuli many more times), the MI would be 0.44 bits. The difference between these two estimates is the bias. Indeed, in this case *#bins* is 15. The total number of stimulus presentation, N, is 40; and therefore the expected bias is 0.27 bits, somewhat larger than the actual observed bias in this case (0.16 bits).

The amount of bias, relative to the estimated information, depends on how densely the joint distribution matrix is sampled. Roughly speaking, this is a problem of *overfitting*. Imagine an experiment in which many different sounds are presented, each of them once, and that the responses are described in so many details that each specific response is seen only once. In this case, in the limited world of the data collected during the experiment, stimuli fully "predict" the responses and vice versa. The decoding can be performed using a lookup table: given a response that actually occurred in the experiment, go back and find the stimulus that evoked it. Thus "information is maximal". However, this result is most probably spurious, in the sense that it does not generalize: for example, another repeat of the same stimulus will probably produce a different response than the one that is in the lookup table. Therefore, the high MI estimated in this case is probably wrong. Accepting the high value of the MI is tantamount to believing that the available data is perfectly representative of all data in all its details, which it is probably not – this is the essence of overfitting.

Estimating the bias using the standard correction given above requires some care, since setting the value of *#bins* is tricky. In experiments, bins of low probability may remain empty because of finite sampling, and therefore the number of non-zero bins in the estimated joint distribution matrix is only a lower bound on *#bins*. On the other hand, assuming that all possible responses are possible to all possible stimuli, hence setting *#bins* to be the total number of bins in the joint-distribution matrix, may also be wrong. For example, it may well be that some stimuli never produce a single spike, while others always produce at least one spike. In this case, some combinations of stimuli and responses are truly empty and will remain so forever. In this case, *#bins* is overestimated. In some respects, this is the situation in the data of Fig. 1 and possibly part of the reason why the estimated bias is somewhat large. Thus, the values of 0 and 1 spikes per stimulus never occur in this sample, and their probabilities, although non-zero, are very small, contributing to the overestimation of the bias. Similarly, large spike counts (>8 spikes/presentation) are extremely unlikely for stimulus 2, also contributing to the relatively large bias correction. Panzeri and Treves (1996) suggested some procedures for better estimating the bias in this context.

There is another possible solution to the problem of overfitting and bias: to reduce the details of the stimulus and response descriptors such that their joint distribution can be better estimated. In standard electrophysiological experiments, a small set of stimuli is used to start with, and therefore the responses should be reduced (e.g. from firing patterns to spike counts). However, the information processing inequality tells us that by doing this, we also reduced information. Thus, we wish to choose the representation in an adaptive manner: as observations accumulate, we increase the amount of details we allow in the representation of the responses. Paninski (2003) showed that by increasing details not too fast as further observations accumulate, such a procedure will converge to the true MI. However, in real experimental situations, the amount of data is fixed, and we cannot follow this scheme to the limit.

We therefore face a conundrum – on the one hand, we cannot fully specify the joint distribution of stimuli and responses in typical experiments because these require too many details, but on the other hand any attempt at reducing the data results in a different problem, where information is already reduced, so that even if we estimate the MI of the reduced problem better, the result is not what we have been looking for to start with.

In practice, one can create a sequence of reasonably well estimated MI values from reduced problems and use this

sequence to generate a high quality final estimate. There are two main approaches to achieve this. In the first approach the MI is calculated for several subproblems and extrapolated to get the expected value at infinite data or infinite resolution. For example, to reduce bias, the MI is calculated for subsets of the data of various sizes. Since the bias decreases theoretically at a rate of $1/N$, it is expected to observe a linear dependence of the observed MI on $1/N$. A linear regression supplies the estimated bias-corrected MI. To get the MI estimate at infinite temporal resolution, the MI may be calculated for several finite temporal resolutions. This is done over a range in which reasonable robust estimation is possible, and these estimates are then extrapolated to infinitely detailed responses. This approach was taken in the so-called direct method (Strong et al., 1998), where estimates of MI were calculated for a series of temporal resolutions, and were linearly extrapolated to infinite temporal precision of spike times.

In the second approach, bias correction can be done on each estimate, resulting in a trade-off: the more reduced joint distribution will have less mutual information, but also smaller bias and less possible over-correction. Thus, standard algorithms may actually show an increase in bias-corrected information, at least initially, as the reduction process proceeds. After producing a series of probable underestimates of the MI, the maximal estimate among all of them should be the best estimator of the true MI. In practice, one has to be careful about the maximization process involved in such procedure, which introduces a bias of its own. A controlled amount of such optimization was used in Nelken et al. (2005), seemingly successfully, at least compared to simulations.

To illustrate these approaches, we use the data of Fig. 4. Suppose we are interested in the phase of the 100 Hz tone. The joint distribution in Fig. 4b was computed using 50 repeats for each stimulus and with 16 bins of 2 ms, covering the range of 0–30 ms. Thus, it contains 128 bins (8 stimulus values and 16 time values for each) and has 400 measurements, slightly more than 3 counts per bin on average. The MI is about 0.79 bits, and the bias (calculated from the number of measurements and number of bins as above) is 0.19 bits, resulting in a bias-corrected MI of about 0.6 bits. In this case, the model is simple enough to allow explicit calculation of the joint distributions: the model MI is 0.62 bits. Thus, the value we get by simple bias correction of a limited-resolution joint distribution is not too far from the 'true' MI, despite the finite sampling and data processing that lie between the model and the estimated value.

With the first approach to bias correction, we should compute the MI for subsets of the data and extrapolate to infinity (in practice, $1/N$ is extrapolated to 0). Fig. 6a shows the plot of average raw MI (without bias correction) against $1/N$. For $N = 10$ trials, the estimated value seems to deviate somewhat from the linear trend defined by the values estimated with larger subsets. Using only the MI of these larger subsets in the linear regression, the estimated

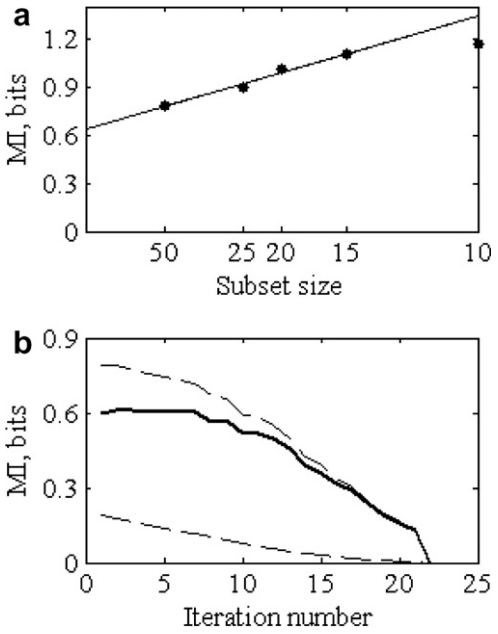

Fig. 6. Bias correction. (a) Estimates of the MI between sine wave phase and responses for the data in Fig. 4a, as a function of the number of trials used. For each number of trials, the largest number of non-overlapping subsets of trials were used (5 subsets for 10 trials, 3 subsets for 15 trials, 2 subsets for 20 and 25 trials, and the whole set for 50 trials). The intercept of the linear regression line with the y-axis is the bias-corrected MI estimate. (b) The sequence of raw MI (upper gray line), bias estimates (lower gray line) and bias-corrected MI (continuous black line) values produced by successive reduction of the joint distribution matrix in Fig. 4b using the algorithm of Nelken et al. (2005).

MI (intercept of the linear regression line with the y-axis) is 0.64 bits, as good an estimate of the true MI as before. Note however that using this method required some judgment for selecting the linear range for the regression, and it assumes that there is enough data so that the bias indeed decreases linearly with $1/N$. If we had only 20 trials per stimulus, we would have available only the first two points of Fig. 6a, and the resulting estimate would have been too large: 0.97 bits.

For the second approach, we use the algorithm developed by Nelken et al. (2005), which successively reduces a joint distribution matrix by joining together adjacent rows or columns. At each step, the row or column with the least marginal probability is selected and joined to the neighbor that has the lower marginal probability. The MI of the reduced matrix and the estimated bias are computed, generating a sequence of bias-corrected MI estimates. Eventually only one column or row remains, stopping the iterations. The maximal bias-corrected value is used as the final MI estimate. Fig. 6b shows the sequence of raw MI values, bias estimates and bias-corrected MI values produced by this algorithm for the same data. While at each iteration of the reduction process both raw MI and bias estimates monotonically decrease, the bias-corrected MI actually increases slightly at the initial stages (although this increase is obviously not significant). The final estimate

of the MI in this case is 0.61 bits, closer to the true data but probably not significantly different from the two other estimates. The main advantage of this approach is that it can be used to get conservative estimates of the MI even with very little data (e.g. Moshitch et al., 2006). In the case of the data used here, with 20 trials per stimulus the estimated MI is 0.63 bits, still close to the model MI. Thus, with small amounts of data, the more conservative second approach seems to perform better.

A number of suggestions have been made for computing MI without explicit calculation of the joint distribution as an intermediate step. These involve the use of decoders (Furukawa and Middlebrooks, 2002; Rolls et al., 1997), computing the MI from a series expansion (which can practically be evaluated only up to the second order, Pola et al., 2003), or using distances between responses (embedded in a high-dimensional metric space) to estimate their density (Victor, 2002). Although these methods do not strictly fall within the context discussed here, we (Nelken et al., 2005) found that in practice, they suffer from the same problems as the matrix-based methods, with the errors being dominated by bias and having a larger mean-squared error than matrix-based methods.

3. Applications to the auditory system

Information-theoretic measures have been applied successfully for problems of auditory coding since the beginning of the current wave of interest in these methods (Rieke et al., 1995). Here I summarize a number of recent studies. This is not intended as an exhaustive review but rather as an illustration of the current practice.

Starting first with a somewhat atypical case, Slee et al. (2005) studied the responses of nucleus laminaris neurons in chick embryos. The stimulus was a non-white Gaussian current. The analysis proceeded in two steps: first a model was fitted to the data, based either on spike-triggered averaging or on covariance decomposition (Brenner et al., 2000), an approach that generalizes and extends spike triggered averaging. In both cases, the procedure generates a model in which linear filter functions are used to reduce the stimulus, and the reduced stimulus undergoes non-linear transformation into firing rate. The issue in this case is whether the simplified model captures all stimulus features that are relevant for spiking, and this was tested using the methods discussed above. The full MI between stimuli and responses was estimated, and so was the MI between the reduced models and the responses. By the information-processing inequality, the second value is always smaller than the first. Better models give MI values that are closer to the full MI. In this study, models captured 60–75% of the full MI. The authors suggest that unexplained MI may have to do with spikes that are very close to each other (<5 ms), in which case spike history influences firing in ways that the reduced models do not capture.

Higher up in the auditory system, Chase and Young (2005) studied the coding of multiple cues for space in cat inferior colliculus (IC). The question they addressed was that of segregation of processing pathways through the IC. Previous studies have suggested that brainstem nuclei which process different spatial cues project to segregated domains in the IC. If so, it could be expected that IC neurons would show mostly sensitivity to the cue that is represented in their dominant input. Chase and Young used stimuli in which various cues (interaural level differences; interaural time differences; and spectral notches) were manipulated independently (therefore mostly working with artificial combinations of cues). This experimental design allowed them to study the way one cue is coded in the presence of variations in other cues, and also to study information interactions between different cues. Their main conclusion is that information interactions are large and are seemingly inconsistent with hard segregation, supporting rather the notion that IC neurons typically integrate information from multiple input streams.

Still in the IC, Escabi et al. (2003) compared the responses to artificial sound ensembles whose spectro-temporal modulations resembled to varying extent those of natural sounds. They found that more naturalistic ensembles evoked higher firing rates and also higher mutual information rates. Interestingly, the information per spike remained about constant, suggesting that the spikes encoded information independently of each other, and that the increase in MI for the naturalistic ensembles was due to the higher firing rates rather than to changes in the way spikes encode stimulus features.

Hsu et al. (2004) analyzed responses of neurons in the midbrain, primary forebrain areas and secondary forebrain areas in a bird, the zebra finch. The basic issue was again that of specialization for natural sound ensembles, and this question was tested by using a hierarchy of ensembles that approximated to varying extent a set of conspecific natural vocalizations. The relative level of MI was used to indicate selectivity. In contrast with other studies reviewed here, the MI was estimated using semi-parametric models: first, a statistical model of the spike train was generated for each stimulus, and then the MI of the models was estimated. They found that the selectivity of neurons to the natural sound ensemble increased at higher auditory stations. As in the paper of Escabi et al. (2003), this increase was not due to better reliability of the individual spikes, but rather to a more extreme distribution of firing rates and higher bandwidth of firing rate modulations in the responses to the more natural sound ensembles. The differences in information/spike for the different sound ensembles in the different stations were not very large, however.

In contrast with other studies reviewed here, Lu and Wang (2004) measured entropies, rather than MI, of spike trains in the auditory cortex of awake marmosets. The purpose was to check the presence of specific spike patterns that are not locked to the stimulus and therefore won't be apparent in the peri-stimulus time histograms. They

analyzed the responses to periodic sounds, for which they demonstrated the presence of two populations, one that locked to specific periods and the second that responded by graded firing rates to different periods but did not show any locking. Lu and Wang were interested in checking whether there are repeating spike patterns even in the responses of the non-locking neurons. Their approach was to jitter spike timing to varying degree and look for increase in the entropy of the spike trains. They could demonstrate such increases for the neurons that locked to periodic sounds, but not to the non-locking neurons, concluding that there are no special firing patterns in the responses of the non-locking neurons.

Middlebrooks and coworkers used decoders, quantified by their transmitted information, to study the coding of space in auditory cortex of anesthetized and awake cats. The use of these methods was initially due to the fact that neurons in auditory cortex of cats tend to respond to a stimuli from large extent of space – a hemifield or even omnidirectionally – but nevertheless their firing patterns may depend on space. Thus, standard measures of spatial selectivity, such as best direction and angular width of the receptive fields, do not make much sense. Middlebrooks called these neurons 'panoramic' (Middlebrooks et al., 1994), and eventually used the transmitted information of a decoder to quantify their coding capabilities. Later, MI was used to study information-bearing elements in the responses as a way of addressing issues of the neural code (Furukawa and Middlebrooks, 2002). Currently, these studies form the best and most extensive study of a single coding task in different auditory cortex fields (Stecker et al., 2005).

To conclude this list, Nelken et al. (2005) used the information processing inequality explicitly as a tool to find candidates for the neural code. They estimated the full information in the spike trains with a number of different computational approach, concluding that the direct method, with a carefully balancing of bias and information loss, is the best. They then demonstrated that two reduced features, spike counts and mean burst latency, do not reach the full information. However, jointly these two variables extracted information levels that were highly similar to the full information. Thus, the two variables jointly may serve as the neural code.

4. Concluding remarks

Information-theoretic tools can serve to study coding problems beyond those surveyed above. Although most studies of the auditory system currently analyze one neuron at a time, some studies of coding interactions between neurons have already appeared (e.g. Chechik et al., 2006; demonstrating high degree of independence between pairs of auditory cortex neurons). The issue of coding interactions was explicitly kept out of this review, but with the increase in the availability of simultaneous recordings it will become as important as that of stimulus coding.

So, is it worthwhile to use information-theoretic measures? The answer is an emphatic yes, given a number of cautionary remarks.

First, the experimental question should justify the use of the MI. In all cases reviewed above, it was difficult to even pose the experimental question in terms of classical measures. Although this is possible, by using measures of reliability and signal-to-noise ratios, the optimality properties of the MI play an important role in these studies. On the other hand, the demonstration of significant stimulus–response associations can usually be done using simpler methods than the MI.

Second, the experiment should be designed with the use of the MI in mind. Thus, using MI entails collecting more data than is usually necessary for estimating just mean rates. It is a bad idea to use MI on sparse data, hoping that its use would miraculously uncover associations that were missed before. As a rough guideline, when analyzing the data, the number of bins in the joint distribution matrix should be smaller than the total number of stimulus presentations.

Finally, the method used to estimate the MI should be carefully evaluated, preferably on surrogate data for which the MI is known and which is as similar as possible to the real data for which the MI is estimated. This step is important because of the difficulties in estimating MI. Although some major advances in our understanding of MI estimation have been made (e.g. Paninski, 2003), many methods that are being used in the literature do not fit the better-understood frameworks.

Acknowledgements

This work was supported by a grant from the Israeli Science Foundation. We thank Yael Bitterman for useful comments on the manuscript.

References

Brenner, N., Bialek, W., de Ruyter van Steveninck, R., 2000. Adaptive rescaling maximizes information transmission. Neuron 26, 695–702.

Chase, S.M., Young, E.D., 2005. Limited segregation of different types of sound localization information among classes of units in the inferior colliculus. J. Neurosci. 25, 7575–7585.

Chechik, G., Anderson, M.J., Bar-Yosef, O., Young, E.D., Tishby, N., Nelken, I., 2006. Reduction of information redundancy in the ascending auditory pathway. Neuron 51, 359–368.

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

Deneve, S., Latham, P.E., Pouget, A., 2001. Efficient computation and cue integration with noisy population codes. Nat. Neurosci. 4, 826–831.

Escabi, M.A., Miller, L.M., Read, H.L., Schreiner, C.E., 2003. Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J. Neurosci. 23, 11489–11504.

Furukawa, S., Middlebrooks, J.C., 2002. Cortical representation of auditory space: information-bearing features of spike patterns. J. Neurophysiol. 87, 1749–1762.

Hsu, A., Woolley, S.M., Fremouw, T.E., Theunissen, F.E., 2004. Modulation power and phase spectrum of natural sounds enhance

neural encoding performed by single auditory neurons. J. Neurosci. 24, 9201–9211.

Lu, T., Wang, X., 2004. Information content of auditory cortical responses to time-varying acoustic stimuli. J. Neurophysiol. 91, 301–313.

Middlebrooks, J.C., Clock, A.E., Xu, L., Green, D.M., 1994. A panoramic code for sound location by cortical neurons [see comments]. Science 264, 842–844.

Moshitch, D., Las, L., Ulanovsky, N., Bar-Yosef, O., Nelken, I., 2006. Responses of neurons in primary auditory cortex (A1) to pure tones in the halothane-anesthetized cat. J. Neurophysiol. 95, 3756–3769.

Nelken, I., Chechik, G., Mrsic-Flogel, T.D., King, A.J., Schnupp, J.W., 2005. Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. J. Comput. Neurosci. 19, 199–221.

Nemenman, I., Bialek, W., de Ruyter van Steveninck, R., 2004. Entropy and information in neural spike trains: progress on the sampling problem. Phys. Rev. E – Stat. Nonlin. Soft Matter Phys. 69, 056111.

Nirenberg, S., Latham, P.E., 2003. Decoding neuronal spike trains: how important are correlations?. Proc. Natl. Acad. Sci. USA 100 7348–7353.

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.

Pola, G., Thiele, A., Hoffmann, K.P., Panzeri, S., 2003. An exact method to quantify the information transmitted by different mechanisms of correlational coding. Network 14, 35–60.

Rieke, F., Bodnar, D.A., Bialek, W., 1995. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. Proc. R. Soc. Lond. B Biol. Sci. 262, 259–265.

Rolls, E.T., Treves, A., Tovee, M.J., 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.

Schneidman, E., Bialek, W., Berry, M.J., 2003. Synergy, redundancy, and independence in population codes. J. Neurosci. 23, 11539–11553.

Sharpee, T., Rust, N.C., Bialek, W., 2004. Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput. 16, 223–250.

Slee, S.J., Higgs, M.H., Fairhall, A.L., Spain, W.J., 2005. Two-dimensional time coding in the auditory brainstem. J. Neurosci. 25, 9978–9988.

Sokal, R.R., Rohlf, F.J., 1981. Biometry, second ed. W.H. Freeman, New York.

Stecker, G.C., Harrington, I.A., Middlebrooks, J.C., 2005. Location coding by opponent neural populations in the auditory cortex. PLoS Biol. 3, e78.

Strong, S.P., de Ruyter van Steveninck, R.R., Bialek, W., Koberle, R., 1998. On the application of information theory to neural spike trains. Pac. Symp. Biocomput., 621–632.

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

Victor, J.D., 2002. Binless strategies for estimation of information from neural data. Phys. Rev. E 66, 51903.