A functional interpretation of global-to-fine refinement in inferior temporal cortex

Lars Schwabe and Klaus Obermayer

Dept. of Electrical Engeneering and Computer Science, TU Berlin
Franklin Str. 28/29, 10587 Berlin, Germany

Abstract

Usually it is assumed that the firing of neurons in the visual cortex represents the presence or absence of a so-called preferred stimulus with the firing of a neuron carrying the same type of information about a flashed but otherwise static stimulus at the beginning, middle and end of a response period which lasts $100 - 200$ ms. Recently, however, it has been found that the instantanous firing rate of neurons in monkey inferior temporal cortex first signals only global information with details of the stimulus being represented only at later times during the reponse period. In this paper we propose a population code model which suggests to interpret this global-to-fine refinement as the signature of communicating stimulus information from inferior temporal cortex to an iterative read-out in, e. g., the prefrontal cortex.

1 Introduction

Understanding how the brain transforms signals from the sensory periphery into more abstract representations which serve as the basis for selecting actions among a set of possible alternatives is a topic of great interest. These transformations are likely to depend on the behavioral context and the animals internal state like, e. g., intentions or expectations. Until now, however, even the transformation of sensory signals in the absence of task-specific modulations is not completely understood. In general, one is interested in *what* neurons in sensory areas are representing about the outside world and *how* this information is represented by the firing neurons. From a theoretical point of view it is hard to say anything general about what aspects of the environment should be represented without referring to a specific sensory modality and a specific environment which makes it mainly an experimental question. Nevertheless, the question of how to represent relevant information is accessible to theoretical studies.

Preprint submitted to Elsevier Science *23rd February 2003*

Previously it has been assumed that the firing of a neuron signals the presence of its so-called preferred stimulus, and that this relation holds for the entire response period. In the context of representing visual information, e. g., it has been assumed that face-selective neurons in the inferior temporal cortex signal the presence of the face to which they are selective at the beginning, middle and end of the response period which lasts a few hundred milliseconds. Recently, however, Sugase and co-workers [1] have found in a passive viewing experiment without any task-demands that the instantaneous firing rate of neurons in the monkey inferior temporal cortex initially represents more global information with the response becoming more selective only at later times. This seems to imply that the temporal modulation of the firing rate carries additional information about the stimulus, and a potential read-out mechanism needs to incorporate this information for successful decoding. Although one may question whether explicit decoding is done at all in the brain, the inferior temporal cortex which projects to prefrontal cortex is at least the most plausible area where a representation might be read out, because one function attributed to the prefrontal cortex is to plan and initiate behavioral responses.

In this paper we set up a simple model of a neuronal population in inferior temporal cortex whose spikes represent a visual stimulus. The spikes are read out in an iterative way by a decoder assumed as a circuit in prefrontal cortex. We show that the dynamically changing stimulus selectivity of the instantaneous firing rate as observed by Sugase and co-workers emerges as an optimal encoding strategy.

Although the actual computations involved when simulating our model (see Fig. 1) may also be implemented with neuronal elements, our model is indeed intended to be a 'black-box model' in the sense of not making explicit the detailed mechanisms underlying the firing rate's temporal variation or how to implement the decoder. Nevertheless, we model at least two important observations. First, we assume the neurons whose spikes serve as the stimulus' representation to be noisy computing devices with their time-varying firing rates describing the instantaneous probability of firing a spike, i. e. we use binary stochastic model neurons. Second, we assume the stimulus-dependent firing rates of neurons which preferentially respond to stimuli within one already learned category to be similar compared with the firing rates of neurons representing stimuli within a different category. We model this by describing the firing rates of neurons using Gaussian tuning functions centered at a preferred stimulus'. Let us now describe the model in greater detail.

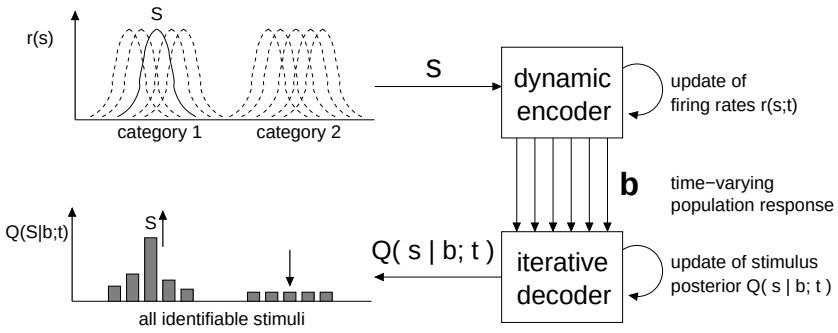

Figure 1. Illustration of the model. A stimulus s is randomly drawn from the set of stimuli within category 1 and category 2 and induces the time-varying firing rates $\mathbf{r}(s;\, t)$ of the neurons within a population of neurons in inferior temporal cortex (the encoder). The also time-varying spike patterns $\mathbf{b}$ are generated based on the time-varying firing rates $\mathbf{r}(s;\, t)$. A downstream population (the decoder) which observes the spike patterns implements an iterative read-out in the sense of updating in each time-interval its 'belief' about which stimulus has caused the observed response. We assume this read-out to express its 'belief' in terms of probabilities and to use the Bayesian inference rule to update them. For the sake of simplicity, we also assume that for decoding the same probabilistic model is used as for the encoding. As more spikes are observed by the read-out, one may expect the posterior belief $Q(s \mid \mathbf{b};\, t)$ for stimulus s which caused the response to increase, whereas the probabilities for the other stimuli decrease.

2 Model

In the following we utilize the technical terms 'encoder' and 'decoder' borrowed from the discipline of communication and information theory. Although our model is certainly applicable to the engeneering problem of ensuring reliable communication over noisy channels, we intent to be specific in the sense of modeling the interconnected areas inferior temporal cortex and prefrontal cortex. Thus, the 'encoder' corresponds to populations of neurons in inferior temporal cortex having inter-areal connections to circuits in prefrontal cortex which should be viewed as the 'decoder'.

Encoder model: We model the responses of a population of N neurons to $M = N$ possible stimuli. With each stimulus we associate a position in a one-dimensional stimulus-space. To the first $M/2$ stimuli belonging to category 1 we assign equally separated positions between $-W-1$ and $-W+1$, whereas to the second $M/2$ stimuli belonging to category 2 we assign equally separated positions between $W - 1$ and $W + 1$. Thus, W determines the separation between the categories in stimulus space. We set $W = 5$ and $N = M = 22$. With each neuron i we also associate a preferred position x_i in stimulus space. Since we use $M = N$ we identified these positions with the corresponding stimulus positions. In other words, each neuron can be identified with one stimulus. The firing rate response of the i-th neuron at time step t after onset

of stimulus s is given by

$$r_i(s;\, t) = R_{\max} \cdot \exp\left(-\frac{(x(s) - x_i)^2}{2\sigma_t^2}\right),$$

where $R_{\max} = 50$ sp/s is the maximal firing rate, $x(s)$ is the position of s in stimulus space, x_i is the preferred position of neuron i in stimulus space and σ_t denotes the possibly time-varying width of this response function. Now, we discretize time into short intervals of duration τ and consider $\tau \cdot r_i(s;\, t)$ as the probability of neuron i firing in response to stimulus s in the t-th time interval, i. e.

$$P\left(b_i^t \mid s;\, t\right) = b_i^t \underbrace{\left(\tau r_i(s;\, t)\right)}_{P\left(b_i^t=1 \mid S=s;\, t\right)} + \left(1 - b_i^t\right) \underbrace{\left(1 - \tau r_i(s;\, t)\right)}_{P\left(b_i^t=0 \mid S=s;\, t\right)}$$

for $b_i^t \in \{0,\, 1\}$ being the response of neuron i in the t-th interval.

Decoder model: We assume that the decoding is done in an iterative way instead of being based on the whole memorized spike pattern beginning with stimulus onset. We model this by assuming the decoder to have an internal state representing at each time step t the multinomial prior probability $P(s;\, t)$ over the set of discrete stimuli. After having observed the population response $\mathbf{b}^t$ in the t-th time interval, this internal state is updated via $P(s;\, t+1) = Q(s \mid \mathbf{b}^t;\, t)$, with

$$Q\left(s \mid \mathbf{b}^t;\, t\right) = \frac{P(\mathbf{b}^t \mid s;\, t) \cdot P(s;\, t)}{\sum_{s'} P(\mathbf{b}^t \mid s';\, t) \cdot P(s';\, t)} \quad \text{and} \quad P\left(\mathbf{b}^t \mid s;\, t\right) = \prod_{i=1}^{N} P\left(b_i^t \mid s;\, t\right).$$

In other words, we assume the decoder to perform an iterative Bayesian estimation of the stimulus using the same model $P(\mathbf{b}^t \mid s;\, t)$ as was used by the encoder to generate the population response $\mathbf{b}^t$.

Objective for encoding/decoding: In order to compare different encoding strategies, one needs an 'objective yardstick'. We set up a reward matrix $\mathbf{R}$ with each R_{sk} representing the reward obtained when decoding stimulus s as k. Since the decoder has at each time-step an internal state representing a multinomial probability over the stimuli, we choose the expected reward

$$E[R;\, t] = E\Big[\quad E[R_{sk}]_{P(k \mid S=s)} \quad\Big]_{P(s)}$$

as a time-dependent measure to compare different coding strategies [1] for different reward matrices. Here, $E[\cdot]$ denotes the expectation operation, $P(s) =$

[1] This measure assumes that the decoding is done by randomly drawing the decoded stimulus based on the decoder's current internal state instead of selecting the stimulus with maximal probability which might be another decoding strategy.

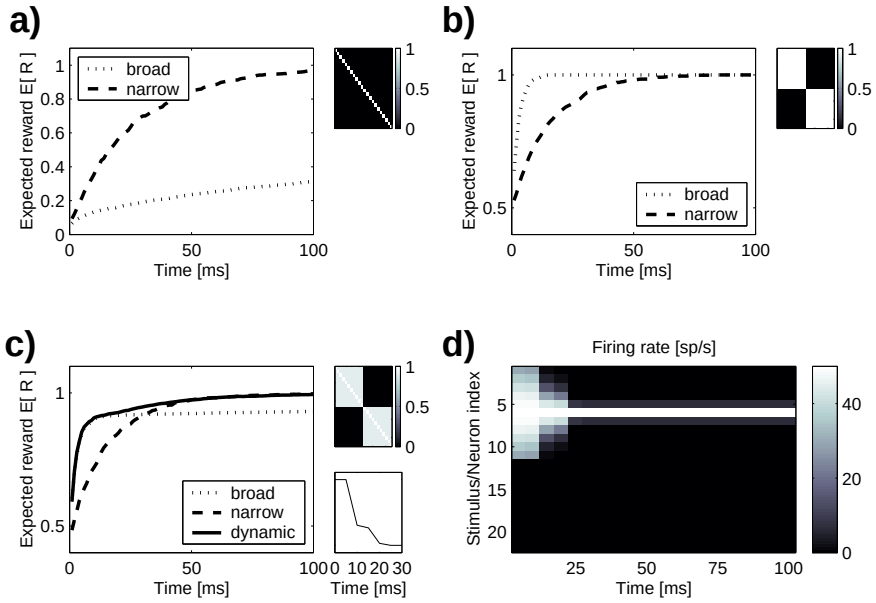

Figure 2. **a)** Expected performance for static encoders in a pure identification task. The top right boxes in Figs. a-c show the corresponding reward matrix (see text). **b)** Expected performance for static encoders in a pure categorization task. **c)** Expected performance for static and dynamic encoders in a combined categorization/identification task. The bottom right box shows how the tuning width of all neurons decreases with time. **d)** Predicted firing rates of the encoder neurons in response to the 6th stimulus.

$1/M$ is the probability[2] of presenting stimulus s, and $P(k \mid S = s; t)$ is the probability of decoding stimulus s as k at time step t, i. e. the expected internal state averaged over the noise in the population response.

3 Results

Predictions for static encoders: Fig. 2a,b show the time-dependent expected reward for two static encoder models and reward matrices. In Fig. 2a we used a reward matrix $R_{sk} = \delta_{sk}$ which yields a reward of 1 only if the stimulus is decoded correctly (correct identification). The dashed line corresponds to an encoder model with $\sigma_t = 0.1$ for all neurons and time-steps, i. e. a narrow tuning curve with the neurons responding almost exclusively to a single stimulus. In contrast, the dotted line corresponds to $\sigma_t = 1$ for all neurons and time-steps with the neurons responding also to other stimuli within a category. One observes that narrow tuning is superior to broad tuning in this case. In contrast, in Fig. 2b we used a reward matrix which yields a reward of 1 even if the stimulus is not decoded correctly as long as the decoded

[2] Note, here $P(s)$ describes the environment's statistics and not an internal state of the decoder, although we set this internal state at $t = 0$ to $1/M$ as well.

stimulus belongs to the same category (correct categorization). In this case, one observes that broad tuning which leads to high activity of many neurons is superior to narrow tuning with very selective responses. These results show that the best encoding strategy may depend strongly on the assumed reward matrix.

Predictions for a dynamic encoder: Given that the reward matrix has substantial influence on the encoding strategy to be utilized, we also explored a third reward matrix which yields a reward of 1 if the stimulus is decoded correctly, a reward of 0.9 if the decoded stimulus belongs to the same category as the encoded stimulus, and no reward if it belongs to the other category. Fig. 2c shows that the curves for the time-dependent rewards cross each other with initially broad tuning being superior to narrow tuning. This can be understood by noting that an encoder with narrow tuning curves almost always leads to successful identification of the stimulus, but only after a delay of approx. 100 ms. The reward obtainable with an encoder having broad tuning curves is limited by the reward given for correct categorizations, but only very seldom the extra reward for correct identification is received. The advantages of both strategies, however, can be combined by changing the tuning dynamically, i. e. to select the optimal σ_t^{op} for a given reward matrix via

$$\sigma_{t+1}^{\mathrm{op}} = \arg\max_{\sigma} \quad E\left[R;\, t,\, \sigma_1^{\mathrm{op}},\, \sigma_2^{\mathrm{op}},\, \ldots,\, \sigma_t^{\mathrm{op}},\, \sigma_{t+1} = \sigma\right].$$

The σ_t^{op} selected according to this procedure are shown in 2c (bottom right). The corresponding expected reward (Fig. 2c, solid line) indeed combines the advantages of both encoding strategies. As an illustration, Fig. 2d shows a population response of this dynamic encoder.

4 Discussion

In summary, we have set up a population code model for communicating information about discrete stimuli by means of time-varying firing rates which are read out by an iterative decoder. This model has been inspired by the findings of Sugase and co-workers [1] who found that the instantanous firing rate of neurons in monkey inferior temporal cortex first signals global information about a stimulus with details being represented only at later times during the response period. Our results suggests that this global-to-fine refinement can be interpreted as the signature of the communication of stimulus information from inferior temporal cortex to an iterative read-out in, e. g., the prefrontal cortex.

Biological relevance: Based on the findings of Sugase and co-workers one can raise a series of questions (see also [2]). For example, why have Sugase

and co-workers not found many neurons which show specific responses at the beginning of the response period corresponding to narrow tuning curves (see dashed lines in Figs. 2a-c) in our model? We have shown that an encoder with initially narrow tuning curves is superior to an encoder with initially broad tuning curves only if fine discrimination (see Fig. 2a) is important at the very beginning. Given that in [1] a passive viewing paradigm was used which excluded any task-specific top-down modulations, one may conclude that a monkey's implicit reward matrix for passive viewing is similar to the combined categorization/identification matrix (see Fig. 2c), because only in this case our model predicts global-to-fine refinement. Since this reward matrix seems to be biologically much more plausible than the identification or categorization matrix, we view our (at least qualitatively) reproduction of the results in [1] as evidence for our model, i. e. in particular for the idea of an iterative read-out according to which the representation is optimized. One could also ask how a read-out can utilize the information present in the temporal modulation of the firing rate. Within our model the decoder does not utilize this information explicitly, but in an implicit way by means of the time-varying internal decoder state. Even with our rather simple model we can make an experimental prediction: Changing the task-demands from passive viewing (or categorization) to identification effects the time-course of the initial reponses in the sense of making the responses more selective which may appear as if the encoder neurons increase their mutual competition.

Relation to other work: In our previous works (see, e. g., [3]) we have already utilized the idea of optimizing a representation for an iterative read-out, but in the context low-level representations in primary visual cortex and without modeling the read-out explicitly. In another theoretical [4] work the idea of an iterative read-out has also been utilized, but the population code representation was assumed to be fixed instead of being dynamically optimized, and in the methodological work [5] an iterative read-out has been developed for decoding experimentally recorded spike-trains. We like to point out that our model focuses on the aspect of communicating already extracted information instead of explaining why and how this information is extracted in the first place like, e. g., approaches which propose to view visual cortex as implementing probabilistic generative models [6,7] with the dynamics of neuronal responses being the signature of inference processes.

Outlook: In at least four aspects the current model needs to be refined. First, one needs to make explicit the mechanisms underlying the time-varying firing rates. Second, one needs to consider also the way by which the stimulus information is extracted in the first place. Third, one needs to explore how task-specificity influences the dynamic representation. Fourth, one needs to develop a biologically plausible architecture for the decoder which allows for making experimental predictions. In the final version of this paper we plan to incorporate the last aspect, i. e. utilizing a set decoder neurons whose firing

rates represent the internal state.

References

[1] Y. Sugase, S. Yamane, S. Ueno, K. Kawano, Global and fine information coded by single neurons in the temporal visual cortex, Nature 400.

[2] M. W. Oram, B. J. Richmond, I see a face - a happy face, Nat Neurosci. 2 (10) (1999) 856–8.

[3] P. Adorjan, L. Schwabe, G. Wenning, K. Obermayer, Rapid adaptation to internal states as a coding strategy in visual cortex?, NeuroReport 13 (3) (2002) 337–342.

[4] J. R. Otterpohl, K. Pawelzik, Adaptive control of expectation as a model for attentional modulation in feature-based visual search, in: Abstracts Society for Neuroscience, 2001 Annual Meeting of the Society for Neuroscience, 2001.

[5] M. C. Wiener, B. J. Richmond, Model based decoding of spike trains, Biosystems 67 (1-3) (2002) 295–300.

[6] R. P. N. Rao, D. H. Ballard, Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects, Nat Neurosci. 2.

[7] G. E. Hinton, P. Dayan, B. J. Frey, R. M. Neal, The 'wake-sleep' algorithm for unsupervised neural networks, Science 268 (5214) (1995) 1158–61.