Adaptive integration in the visual cortex by depressing recurrent cortical circuits

LETTER **Communicated by Matteo Carandini**

Adaptive Integration in the Visual Cortex by Depressing Recurrent Cortical Circuits

Mark C. W. van Rossum
mvanross@inf.ed.ac.uk
Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, EH1 2QL, U.K.

Matthijs A. A. van der Meer
mvdm@umn.edu
Doctoral Training Center Neuroinformatics, School of Informatics, University of Edinburgh, Edinburgh, EH1 2QL, U.K.

Dengke Xiao
dx@st-andrews.ac.uk
Mike W. Oram
mwo@st-andrews.ac.uk
School of Psychology, St. Andrews University, St. Andrews, Fife, KY 16 9JU, U.K.

Neurons in the visual cortex receive a large amount of input from recurrent connections, yet the functional role of these connections remains unclear. Here we explore networks with strong recurrence in a computational model and show that short-term depression of the synapses in the recurrent loops implements an adaptive filter. This allows the visual system to respond reliably to deteriorated stimuli yet quickly to high-quality stimuli. For low-contrast stimuli, the model predicts long response latencies, whereas latencies are short for high-contrast stimuli. This is consistent with physiological data showing that in higher visual areas, latencies can increase more than 100 ms at low contrast compared to high contrast. Moreover, when presented with briefly flashed stimuli, the model predicts stereotypical responses that outlast the stimulus, again consistent with physiological findings. The adaptive properties of the model suggest that the abundant recurrent connections found in visual cortex serve to adapt the network's time constant in accordance with the stimulus and normalizes neuronal signals such that processing is as fast as possible while maintaining reliability.

1 Introduction

Input to the visual system is extremely variable, and the visual signal from a given object can vary in properties such as size, position, and orientation. It has long been realized that one of the roles of the visual system is to

remove these stimulus variations. In a layered network, such invariant object representations can be obtained by combining responses of neurons with different receptive fields at various stages in the processing stream (Riesenhuber & Poggio, 1999). However, stimulus contrast and stimulus duration typically also vary. Invariance to these changes cannot be obtained by pooling neurons, but it can be achieved using temporal integration. For instance, a low-contrast stimulus can be integrated longer to maintain a signal-to-noise ratio similar to that of a high-contrast stimulus. This increase in the temporal integration at low contrast is reflected in the latency of the neural responses. As the stimulus contrast is lowered, response latencies increase. This already happens in retina (Shapley & Victor, 1978), but also occurs in the lateral geniculate nucleus (LGN; Lee, Elepfandt, & Virsu, 1981), V1 (Albrecht & Hamilton, 1982; Dean & Tolhurst, 1986; Carandini & Heeger, 1994; Albrecht, 1995; Saul, 1995; Gawne, Kjaer, & Richmond, 1996; Bair, Cavanaugh, Smith, & Movshon, 2002; Albrecht, Geisler, Frazer, & Crane, 2002), area MT (Raiguel, Xiao, Marcar, & Orban, 1999), and the anterior superior temporal sulcus (STSa) (Oram, Xiao, Dritschel, & Payne, 2002). The impact of contrast on response latencies becomes progressively larger in higher areas: as contrast is lowered, the latency of V1 responses increases from about 40 ms to 75 ms (Gawne et al., 1996), but in area STSa, response latency increases from about 90 ms to 225 ms (Oram et al., 2002), arguing for additional latencies incurred at each processing stage.

Small contrast-dependent changes in response latency, such as those observed in V1, can be accounted for in many ways (Bugmann & Taylor, 1993; Carandini & Heeger, 1994; Bair et al., 2002). A number of previous models have included synaptic depression to explain temporal V1 response properties (Chance, Nelson, & Abbott, 1998; Carandini, Heeger, & Senn, 2002; Kayser, Priebe, & Miller, 2001; see also Loebel & Tsodyks, 2002), but typically have feedforward connectivity only. This is problematic in the light of evidence that synaptic depression in the drive from LGN to V1 is limited, while polysynaptic connections via other V1 cells depress strongly (Boudreau & Ferster, 2005). Furthermore, the large latency changes in higher areas are inconsistent with pure feedforward models of visual processing. In models of spiking feedforward networks with realistic noise, latencies are short and only weakly dependent on firing rates or stimulus contrast (Knight, 1972; Treves, 1993; Gerstner, 2000; van Rossum, Turrigiano, & Nelson, 2002).

In addition to the contrast manipulation, we also consider manipulation of the stimulus duration. It has been observed that when a stimulus is briefly flashed, the response significantly outlasts the stimulus, such that both the response duration and response amplitude in higher visual areas depend only weakly on the precise stimulus duration (Rolls & Tovee, 1994; Keysers, Xiao, Földiák, & Perrett, 2005). This is also incompatible with linear feedforward models.

We show that both the contrast-dependent latencies and the invariance of response to brief stimuli are reproduced in a network model that includes

two physiological observations: strong recurrent excitatory connections known to be abundant in cortex (Douglas, Koch, Mahowald, Martin, & Suarez, 1995) and short-term synaptic depression of these connections (Thomson & West, 1993; Markram & Tsodyks, 1996; Varela et al., 1997). The abundant recurrent connections thus normalize signals by adaptively adjusting the network's gain and time constant.

2 Methods

2.1 Model Definition. We simulate a network model of nodes that are characterized by their firing rates, schematically depicted in Figure 1c. Each node can be thought of as representing the average firing rate of a small population of neurons, such as a microcolumn. In the presence of asynchronous noisy background activity, the dynamics of the population is not limited by the (slow) membrane time constant, but instead the population firing rate responds almost instantaneously to changes in the input current (Knight, 1972; Treves, 1993; Gerstner, 2000; van Rossum et al., 2002). As a result the dynamics is largely determined by the synaptic time course. We therefore model the network as follows (see Dayan & Abbott, 2002, for a discussion of this approximation). The net current $I(t)$ received by a node is described with

$$\tau \frac{dI(t)}{dt} = -I(t) + I_{input}(t) + g P_{rel}(t) r(t), \tag{2.1}$$

where the time constant τ (5 ms) determines the dynamics of the node and reflects the time constant of fast excitatory transmission. The right-hand side contains three terms: a decay term $-I(t)$; an input I_{input} term, which for the first node equals the stimulus and for subsequent nodes equals the synaptic input from the preceding node; and the right-most term, which is the recurrent feedback, indicated by the loops in Figure 1c. The recurrent feedback is subject to short-term synaptic depression and is given by the product of the synaptic release probability $P_{rel}(t)$, the firing rate of the node itself $r(t)$, and the recurrent gain g, which is set to $g = 1$. Note that without depression, a strong recurrent gain ($g \geq 1$) could lead to diverging activity; however, with depressing synapses, a high recurrent gain does not pose a problem.

The synaptic release probability P_{rel} incorporates short-term synaptic depression. The dynamics of the release probability is modeled as a first-order equation. Under the assumption of Poisson firing, the release probability obeys (Tsodyks, Pawelzik, & Markram, 1998)

$$\tau_{\mathrm{depr}} \frac{dP_{\mathrm{rel}}(t)}{dt} = P_0 - [1 + \tau_{\mathrm{depr}} r(t)(1 - f)] P_{\mathrm{rel}}(t), \tag{2.2}$$

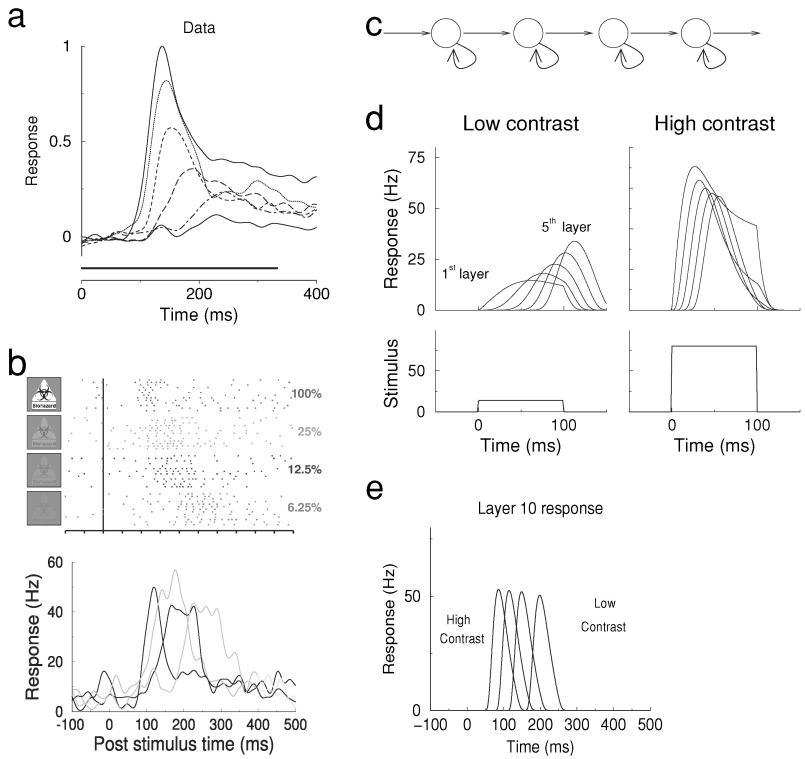

Figure 1: Neural responses in higher visual areas to stimuli presented at different contrasts: data and model. (a) Average normalized responses of 47 STSa neurons measured in response to preferred stimuli of 100%, 75%, 50%, 25%, 12.5%, and 6.25% contrast (top to bottom traces) with 333 ms stimulus duration (bar). (b) Example of responses of an STSa neuron to a preferred stimulus (a biohazard sign) at different contrasts. Rastergrams and corresponding spike density functions (SD = 10 ms) of the responses to multiple presentations of the effective stimulus at different contrasts (100%, 25%, 12.5%, and 6.25%). The response latency increases from 90 ms to 200 ms as the stimulus contrast is decreased. (c) Rate-based network to explain the latency increases of visual signal propagation. In the network, nodes are connected through excitatory connections and receive input from both the previous layer and recurrent excitation. Both the feedforward and recurrent connections are subject to short-term synaptic depression. (d) The activity of the model network in response to a low-contrast (left) and high-contrast (right) step stimulus. The activity is shown in five subsequent layers of the network. For low-contrast stimuli, the latency is substantially longer than for high-contrast stimuli. (e) Model responses in layer 10 to step stimuli of four contrast levels. In higher layers, the responses become stereotypical: duration and amplitude are independent of the input contrast and duration, but the latency varies.

where the parameter $\tau_{\text{depr}} = 500\,\text{ms}$ describes how quickly the synapse recovers toward its default release probability, given by $P_0 = 1$. The depression factor $f = 0.8$ describes how much each spike depresses the release probability ($P_{rel} \rightarrow f P_{rel}$). These parameter values were taken from the literature as they give an accurate description of depression in visual cortical slices (Abbott, Varela, Sen, & Nelson, 1997; Varela et al., 1997) and did not require tuning.

Finally, the firing rate is modeled as an instantaneous function of the current, $r(t) = h(I(t))$, with

$$h(I) = \frac{\kappa \log[\cosh([I]_+/\kappa)]}{1 + \tau_{\text{refr}} \kappa \log[\cosh([I]_+/\kappa)]}.$$

The function $h(I)$, referred to as the F/I curve, is a sigmoid combining two effects. First, it implements a weakly expansive nonlinearity for low firing rates, given by the parameter $\kappa = 5$ spikes/s that characterizes above which frequency the F/I curve becomes approximately linear. Second, the F/I curve is saturating for high rates, determined by the refractory time $\tau_{\text{refr}} = 0.002$s (see Figure 6a). A simple threshold linear F/I curve, $h(I) = [I]_+ = \max(I, 0)$, yields similar behavior but has less sharp transients (see section 3.6).

2.1.1 Feedforward Connections. Higher layers in the network receive their input from the previous layer. Like the recurrent connections, the synapses of these feedforward connections are also subject to synaptic depression; however, the influence of this depression on the latency is weak (see section 3.2). The input current to layer $i + 1$ in equation 2.1 is given by $I^{i+1}_{input}(t) = g_{\text{ff}} P^i_{\text{rel}}(t) r^i(t)$, where $P^i_{\text{rel}}(t)$ and $r^i(t)$ are release probability and the firing rate of the previous layer. An identical feedforward gain g_{ff} is used between all layers. This feedforward gain is adjusted such that a sustained 50 Hz stimulus in the input layer evokes a response in the output layer with a peak rate of 50 Hz. This was done for each model variant independently. For the recurrent networks with depression, this gain is about 0.5. Thus any node receives about twice as much recurrent as feedforward input.

2.1.2 Population Coding Network. For the study of population codes, we use a network with $N = 20$ neurons per layer with wraparound boundaries to eliminate edge effects. The stimuli to the input layer are step stimuli centered around node $k = 10$. The spatial profile of the stimulus is a rectified cosine ("bump"), and the input to node i is $I_{stim,i} = A[\cos(2\pi(i - k)/N)]_+$, where A is the amplitude of the stimulus. The layers are connected to the subsequent layer with an excitatory center and inhibitory surround given by $w^{\text{ff}}_{ik} = g_{\text{ff}} \cos(2\pi(i - k)/N)$, where i and k denote the lateral position of

the nodes in the layers. The feedforward gain g_{ff} is again determined by calibrating the peak responses.

The lateral connectivity matrix equals $w_{ij}^{\mathrm{lat}} = g(1 - \delta_{ij})\cos(2\pi(i - j)/N)$. To demonstrate that self-excitation is not essential for the model, the lateral connectivity matrix explicitly excludes self-excitation. The parameter g is the strength of the lateral connections and is comparable to the recurrent gain g above; its value was determined by matching the latencies of the single-node network to the latencies in the population network. In summary, in the population coding network, each node obeys $\tau \frac{dI(t)}{dt} = -I(t) + I_{input}(t)$, where the input current for a node i is $I_{input}^{i}(t) = \sum_j w_{ij}^{\mathrm{lat}} P_{\mathrm{rel}}^{j}(t) r^j(t) + \sum_k w_{ik}^{\mathrm{ff}} P_{\mathrm{rel}}^{k}(t) r^k(t)$, where j loops over all nodes in the layer of i, while k loops over all nodes in the preceding layer. The synaptic depression parameters are the same as in the single-node case. Both excitatory and inhibitory ($w < 0$) connections depress. We obtained comparable results when inhibitory connections ($w < 0$) were nondepressing.

Axonal and dendritic propagation delays are not included in the model. Feedforward propagation delays add trivially a contrast-independent latency. The delays in the recurrent connections can be substantial due to the lack of myelination in the horizontal connections (Hirsch & Gilbert, 1996; Bringuier, Chavane, Glaeser, & Frégnac, 1999). Including a fixed 10 ms delay in the recurrent connections leads to minor additional latency but does not substantially increase the contrast dependence of the latency.

At the start of the simulation, the synapses are fully recovered, and the firing rate is zero. For compactness, the input and the current are expressed in the same units as the firing rate. Alternatively, constants can be introduced in the F/I curve and the various gains to match the dimensions. This does not change the model.

2.2 Neurophysiological Methods. The experimental protocols have been described before (Oram et al., 2002). Briefly, extracellular single-unit recordings were made using standard techniques from the upper and lower banks of the anterior part of the superior temporal sulcus (STSa) and the inferior temporal cortex (IT) of two monkeys (*Macaca mulatta*) performing a visual fixation task. The subject received a drop of fruit juice reward every 500 ms of fixation ($\pm 3^{\circ}$) while static stimuli ($10^{\circ} \times 12.5^{\circ}$) were displayed. During initial screening, images of different perspective views of monkey and human head, animals, fractal patterns, natural scenes, and everyday objects were presented for 110 ms.Visual inspection of online rasters and the poststimulus time histograms (PSTH) of each stimulus was used to select effective (preferred) and noneffective (nonpreferred) stimuli.

To measure the effect of contrast on the response (see Figures 1 and 2), gray-scale versions of preferred and nonpreferred stimuli were presented for 333 ms followed by a 333 ms interstimulus interval. The different

stimuli and different contrast levels were present in random order. The 100% Michelson contrast $(L_{max} - L_{min})/(L_{max} + L_{min})$ was formed by normalizing the foreground pixel values such that they occupied the monitor's full luminance range after adjusting the initial gray-scale image to have 50% luminance. Other contrast versions (75%, 50%, 25%, 12.5%, and 6.25%) were achieved by systematically varying the width of the distribution of the foreground pixel values of the 100% contrast version while maintaining the average foreground luminance. All manipulations were performed after correcting for the measured gamma function of the display monitor.

2.2.1 Data Analysis. Spike density functions were computed by smoothing a 1 ms bin width peristimulus time histogram with a gaussian filter (SD = 10 ms) for each stimulus at each contrast. Response magnitude was taken as the average firing rate in the 333 ms following response latency. Population-averaged responses were generated by normalizing the spike density function of each cell to the most effective stimulus by setting the average of the 200 ms prior to stimulus onset to 0 and the peak of the spike density function to 1, average across neurons, and renormalizing to the range 0 . . . 1 (Oram & Perrett, 1992).

The latency was extracted at the point at which the activity exceeded the baseline activity (estimated from 200 ms before stimulus onset) by 3 standard deviations for at least 20 ms. The latency was accepted only if the activity of the neuron in the 100 ms following the estimate was significantly ($p < 0.05$) above the baseline activity (paired t-test). We termed this the SD method. The latency was also calculated based on the time at which the spike density function reached half-maximum. The half-maximum estimate of latency is sensitive to random fluctuations in the ongoing activity when the response is small (responses to less preferred or low-contrast stimuli), yielding unreliable latency estimates. Given that we focus here on response latencies at low contrast, we present the data using the statistical SD method.

The model was noise free in most cases, so the SD method was not appropriate, and we used the half-maximum latency. The latencies obtained from the neurophysiological data using either method were highly correlated ($r = 0.87$), which is expected from the steep onset of the responses even at low contrast (see section 3.6).

3 Results

This study combines computational models with recordings from anterior inferotemporal cortex (AIT) and STSa. We first present data showing the contrast dependence of latencies in higher visual areas and introduce a model with strong recurrent connections subject to synaptic depression that explains these data. Next, we show that in the model and the data,

the latency is only weakly determined by the firing rate and stimulus preference. Furthermore, the data and model show similar responses when stimuli are briefly flashed. Finally, we discuss the signal processing and gain regulation in the model.

3.1 Contrast-Dependent Latencies. We first study how response latency in higher visual areas depends on contrast. In Figure 1a the average response of 47 neurons recorded in area STSa is plotted in response to preferred visual stimuli presented at different contrasts. The data show large changes in the latency: the average latency ranges from 90 ms for the highest-contrast stimuli to 216 ms for the lowest-contrast stimuli, while in some neurons, the latency difference across the same contrast range can be in excess of 300 ms (see also Oram et al., 2002). The response of a single neuron in area STSa of four different contrasts is shown in Figure 1b. For this neuron, the response latency at 100% contrast was 89 ms, which increased to 190 ms at 6.25% contrast.

It is noteworthy that at low contrast, the population-averaged responses show less amplitude normalization and a smoother onset of the responses than the single-neuron example. This is partly because the different cells included in the population average have different sensitivity of response latency to contrast. Thus, as the contrast is decreased, the heterogeneous latencies average to a temporally smeared response. However, the onset for a given cell remains steep, with the average time to rise from detected latency to half peak being 8.1 ± 1.1 ms at 100% contrast and 10.2 ± 3.2 at 6.25% contrast, corresponding to 2.6 ± 0.3 and 1.9 ± 0.6 Hz/ms, respectively. Thus, although reducing stimulus contrast increases response latency, the onset remains sharp.

To examine possible mechanisms underlying the change in response latency at low contrast, we study layered networks with the architecture shown in Figure 1c. First we consider a network with just one node per layer, where each node represents a group of neurons with similar receptive fields. The layers are abstract and do not correspond to the anatomical lamina in the cortex. The model solely propagates signal and is of course by no means a full model of the visual cortex. Its purpose is to study the effect of depressing synapses and recurrent connections on the dynamics of signal propagation in the visual system. Nevertheless, networks of this structure can be extended to perform computations (van Rossum & Renart, 2004; Vogels & Abbott, 2005). Based on the known anatomy and physiology, each node receives both feedforward input and strong recurrent excitation (Douglas et al., 1995). These recurrent connections should not be considered all-to-all on a single-neuron level but rather reflect the average strength of the recurrent connections in a group of neurons. Crucially the recurrent connections in the model are subject to short-term synaptic depression, as observed in cortex (Thomson & West, 1993; Markram & Tsodyks, 1996; Varela et al., 1997). Short-term synaptic depression means that the synaptic

response becomes weaker on repeated stimulation, while after a period of rest, the original strength is restored. We took the parameters of the synaptic depression from the literature (see section 2) rather than introducing extra degrees of freedom and fitting them to match the responses.

We first study the propagation of step stimuli of varying contrast through the model network. Stimulus contrast was assumed to be coded in the strength of the model's (retinal) input. Figure 1d shows the response in various layers to stimuli of high and low contrast. Already in the first layer, the response to a low-contrast stimulus has a slower rise than to a high-contrast stimulus. In subsequent layers, additional latency is added at low contrast. In Figure 1e, we show the model response in layer 10 for four contrast levels. The latency increases substantially at low contrasts. Note that some 100 ms after onset, the responses in the data have a sustained portion that is not present in the model. In the model, the response in the higher layers is of limited duration (about 60 ms, basically until the synapses are depressed) even when the stimulus persists. Below, we explore possible explanations for this sustained response.

Next, we directly compared the latencies of the model to the physiology. For the physiological data, the latency at a given contrast was calculated for each neuron and averaged across the population, Figure 2a. For comparison we also show response latencies recorded in area V1 (data from Wiener, Oram, Liu, & Richmond, 2001; Oram, Wiener, Lestienne, & Richmond, 1999). The average response latency in area STSa increases by 33 ± 3 ms for each halving of stimulus contrast, which is significantly greater than in V1, where this was 8 ± 0.8 ms ($F[1,7] = 56.8$, $p < 0.0005$). Thus the majority of the latency change is not of retinal or V1 origin; instead it suggests that each cortical processing area adds latency at low contrast.

The model's latency is plotted as a function of contrast for layers 1, 5, and 10 in Figure 2b (again for a model with one node per layer). To express stimulus rate as a contrast, we used an inverse Naka-Rushton equation, $c = c_{50}(\frac{r/r_{\max}}{1-r/r_{\max}})^{1/n}$, with parameters $c_{50} = 0.5$, $n = 1.6$, corresponding to LGN inputs (Sclar, Maunsell, & Lennie, 1990) and $r_{\max} = 140$ Hz. The minimum latency occurs with the high-contrast stimuli and is approximately equal to the number of layers crossed times the synaptic time constant. In comparing data to model, one should take into account that retinal and propagation delays lead to a latency in layer 4 of V1 of some 50 ms (Maunsell & Gibson, 1992; Schmolesky et al., 1998). It has been estimated that between retina and AIT/STSa, at least 10 synapses must be traversed (Gautrais & Thorpe, 1998; Oram & Perrett, 1992). With this in mind, the latency in the model in layer 10 is comparable to the latency in area STSa for the parameters used. Thus, using realistic parameters and a reasonable number of layers, the latencies in the model are comparable to the neurophysiological data. In the model, the curves are steeper near low contrast than is observed in the data; this is due to the sharper input-output nonlinearity in the model (below).

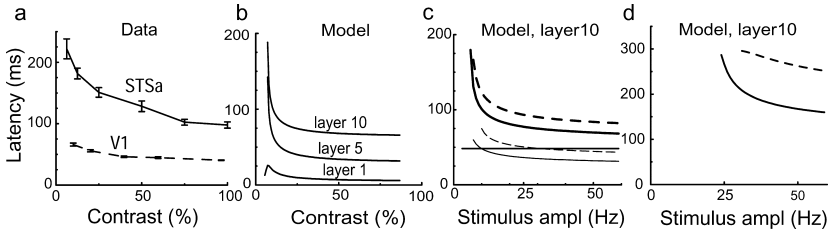

Figure 2: Latencies versus contrast, data and model. (a) Average response latency as a function of contrast measured in 19 V1 neurons in response to Walsh patterns (dashed curve) (data from Wiener et al., 2001; Oram et al., 1999) and in 18 STS neurons (in response to objects) for which latency estimates were available at all six contrast levels (solid curve). Error bars denote standard error. (b) The latency in the network model with depressing recurrent connections as a function of stimulus contrast. The latency is plotted for the first (bottom curve), the fifth, and the tenth layers (top curve). For the weakest stimuli, no latency is plotted in the deeper layers, because the stimulus fails to propagate deeply into the network. (c) Model latencies in layer 10 versus stimulus amplitude for various model variants. The full model with recurrent connections and depression of all synapses has large latency differences (thick solid curve). If the feedforward connections are not depressing, latencies are slightly longer but comparable (thick dashed curve). Without recurrent connections, the maximal latency and its contrast dependence is much smaller (thin solid curve) also when the feedforward connections are not depressing (thin dashed curve). A linear network without depression has a constant latency (straight line). (d) Model latencies in layer 10 versus stimulus amplitude for a model in which the recurrent connections are not depressing. If the feedforward connections are depressing, long-contrast-dependent latencies result (dashed curve). Even longer latencies result when the feedforward connections are not depressing either (solid curve).

3.2 Contributions to Latency Changes. Next, we explored the different contributions to the latency in the model. In Figure 2c, layer 10 latency is plotted versus contrast for a range of model conditions. In a feedforward network, the nonlinearity of the F/I curve by itself leads to a weak contrast-dependent latency (thin dashed curve). In this case, each node filters and (smoothly) thresholds the signal; such models have been used to explain V1 latencies (Bair et al., 2002). Adding depression in the feedforward connections reduces the latency somewhat (thin solid curve), as it reduces the late part of the response. Response latency is, however, much longer with depressing recurrent connections (thick curves) than without recurrence (thin curves). In the presence of the depressing recurrence, depression of feedforward connections again has a small effect on the latency (thick dashed curve).

When the F/I curve is linear ($r = I$) and the synapses are nondepressing, the behavior of the network can easily be studied analytically. In that case, if recurrent feedback is absent, the latency would be proportional to τ times the number of layers crossed, where τ is the synaptic time constant, Figure 2c (straight line). Hence, the latency would be short and independent of contrast. In a linear network with recurrent feedback with strength g (but without synaptic depression), one has

$$\frac{\tau}{1-g}\frac{dr(t)}{dt} = -r(t) + \frac{1}{1-g} I_{input}(t),$$

from which one sees that the gain of each layer is proportional to $1/(1-g)$ and the latency increases to $\tau/(1-g)$, but is still independent of contrast (Douglas et al., 1995). When the synapses are depressing, the system consists of two coupled differential equations per layer, equations 2.1 and 2.2, which complicates matters considerably. In the appendix, we show how the latency can be approximated in that case.

Finally, we consider the case where the recurrent connections are not depressing, Figure 2d. The high recurrent gain $g = 1$ is in this case pathological, so we slightly reduced it to $g = 0.99$. As expected from the above arguments, the latencies are very long. As above, the nonlinearity of the F/I curve and the feedforward depression still lead to contrast-dependent latencies, although the ratio of maximal and minimal latency is again smaller, Figure 2d (solid curve). This seems perhaps an interesting alternative to obtain contrast-dependent latencies. However, experimental evidence does not support such a picture, and if anything, it suggests the opposite (Boudreau & Ferster, 2005). If the feedforward connections are also not depressing, very long contrast-independent latencies result, Figure 2d (dashed curve). Furthermore, as can be inferred from the figure, the minimal contrast required to propagate through the network is higher. This is a consequence of the very high recurrent gain in this case. In order to prevent activity levels that are too high, the feedforward drive g_{ff} is set much lower than in the other model variants (see section 2 for the tuning procedure); this goes at the expense of the low-contrast responses.

In summary, the long latencies observed physiologically are found in the model with depressing recurrence, although factors such as the nonlinearity of the F/I curve and feedforward depression can contribute to the contrast-dependent latency as well. The mechanism is as follows. When the contrast is low, the total input to a node is dominated by the recurrent input, which effectively slows the dynamics (and increases the gain). When presented with a high-contrast stimulus, the recurrent connections are rapidly depressed out, leaving a quick response known from feedforward networks.

3.3 Contrast, not Stimulus Preference, Determines the Latency. A possible interpretation of the above data could be that latency is simply determined by the firing rate of each node. However, it has been observed that response latency in V1 is determined by stimulus contrast, but only weakly, by stimulus preference (Carandini & Heeger, 1994; Albrecht, 1995; Gawne et al., 1996). Comparable to the observations in V1, in STSa and IT a nonpreferred high-contrast stimulus also yields a small response but with a short latency. This is illustrated in Figure 3a (see also Oram et al., 2002; Oram & Perrett, 1992).

To examine the dependence of response latency on response magnitude and stimulus contrast in the model, we implement a population coding network, Figure 3b. Instead of having just one node per layer, each layer in the model now contains an array of 20 neurons. The recurrence was implemented in a lateral connectivity matrix with a center-surround layout (see section 2). Subsequent layers are connected to each other with a weight matrix that implements an excitatory center and inhibitory surround. The stimulus preference for a given node can be changed by placing the stimulus at different locations. Equivalently, we fix the stimulus position and study the response across nodes.

We determined the latency and response amplitude in the population coding network for both a low- and high-contrast stimulus, Figure 3c (left and right). In the network with depressing recurrence, the latency is again strongly contrast dependent, Figure 3c, thick lines. This demonstrates that in population coding networks, depressing connections between neighboring nodes give rise to contrast-dependent latencies. However, it can also be observed that for a given contrast, the latencies in a given layer are very similar. In particular, the activity of central nodes at low contrast (see Figure 3c, lower left) is higher than the activity of the edge nodes at high contrast (see Figure 3c, lower right), yet the latency at low contrast is about twofold longer. In other words, the contrast affects the latency more than the response amplitude of the particular node.

Next we tested a simplified network in which the synapses are not depressing and recurrent connections are absent. As was shown above when the F/I curve is nonlinear, a small contrast-dependent latency remains, Figure 2c. In the population coding network, the latency is again weakly contrast dependent, Figure 3c, thin lines. In this case, one might perhaps have expected a strong coupling between latency and firing rate, but interestingly, for a given stimulus contrast, the latencies within a layer are again quite similar. Although the longer latencies occur for nodes for which the stimulus is less preferred, both contrast and firing rate determine the latency.

In response to high-contrast, nonpreferred stimuli, the latency is short for two reasons. First, unlike with low-contrast stimuli, the latency does not accumulate across layers because nodes with low activation receive a short latency input, mainly driven by nodes in the previous layer with a

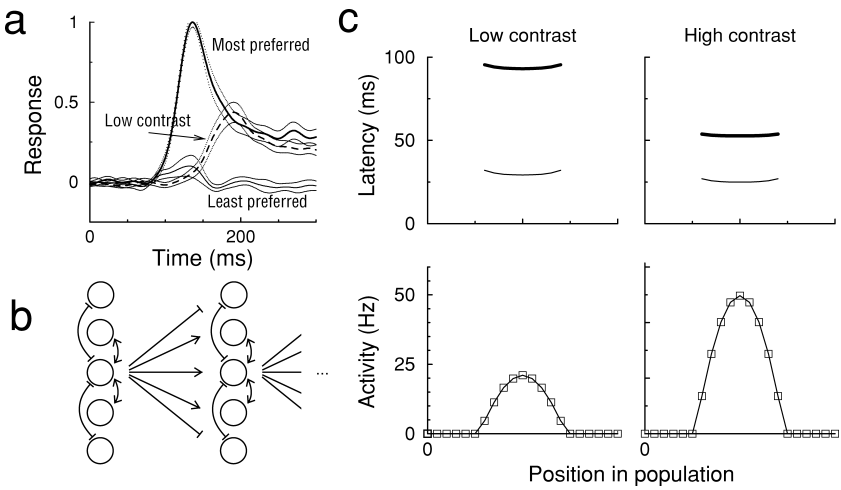

Figure 3: Contrast-dependent latencies in a population coding network. (a) Averaged responses of recorded neurons in area STS to most preferred stimuli (solid curve) and least effective stimuli (thin curve), all presented at high contrast. The least effective stimuli lead to a small response, but with a short latency. For comparison, the response to the preferred stimulus at 25% contrast is also shown (dashed curve). (b) Population coding network architecture in which each layer has 20 nodes. The layers are connected to each other with a center-surround profile. The sharp arrows denote excitatory connections, the blunt arrows inhibitory ones. For clarity, only the connections from the middle nodes are shown. (c) Latencies (top, thick curves) and peak responses (bottom) in the fifth layer of the population coding network. Response to a low-contrast (left) and a high-contrast (right) bump stimulus in the input layer. The latency to high-contrast stimuli is short and similar across nodes, even for nodes on the edges, which have a low firing rate. At low contrast, the latencies are long and again similar across nodes. Note that the most active nodes at low contrast have latencies that are substantially longer than the latencies of weakly active nodes to high-contrast stimuli, indicating that the latency is mainly determined by contrast rather than by firing rate. The thin curves indicate latencies in the model variant without synaptic depression and recurrence (response amplitudes were matched to be identical).

high activity. Second, in the depressing network, the strongest lateral input comes from the nodes with the highest activation. The high activation means these synapses depress quickly, shortening the latency.

Following Gawne et al. (1996), we examined the extent to which response amplitude and latency varied with stimulus identity and stimulus contrast. For recorded cells tested with stimuli that elicited significantly different mean spike counts (ANOVA, $p < 0.05$), stimulus contrast accounted for

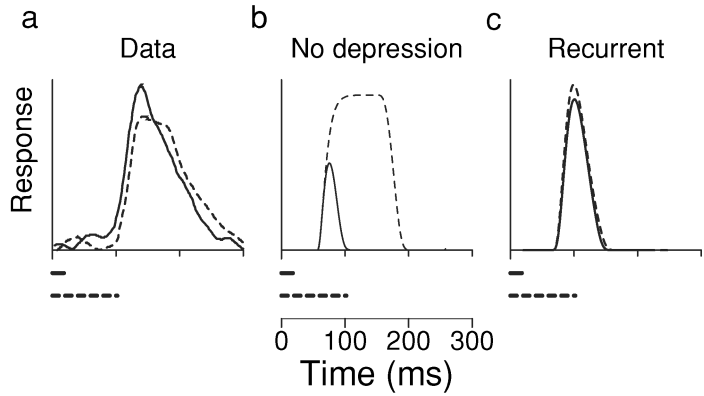

Figure 4: Responses to brief stimuli. (a) Average responses of recorded neurons in area STS to stimuli presented for 18 ms (solid) and for 102 ms (dashed). The responses are almost identical. Preferred stimuli were randomly interleaved with nonpreferred stimuli; only responses to the preferred stimulus were included in the average. (Data from Keysers et al., 2005.) (b) Model responses in a network without synaptic depression and recurrence. The activity in the fifth layer in response to the brief (18 ms) stimulus and prolonged (102 ms) stimulus. In contrast to the data, the response amplitude and duration clearly reflect the stimulus duration. (c) In the depressing recurrent network, the response in the deeper layers becomes independent of stimulus duration. The brief and prolonged stimuli were in all cases presented at identical high contrast; brief, low-contrast stimuli do not propagate through the network.

67 $\pm$ 7% of the variability of response latency and only 33 $\pm$ 3% of the variability in spike count. Conversely, stimuli identity accounted for 69 $\pm$ 6% of the variability in spike count and only 20 $\pm$ 5% of the variability on response latency. Thus, in areas STSa and IT, stimulus contrast is encoded mostly by response latency, whereas stimulus identity is encoded mostly by response magnitude; the same reversal was observed in V1 (Gawne et al., 1996). The same is observed in the model where the contrast accounted for 92% of the variability in the latency and 14% of the response amplitude variation, while the stimulus identity (i.e., position) contributed 80% to the response amplitude variation and 0.1% to the latency (fifth-layer response, linear model fit using cosine of angular stimulus location, and logarithmic stimulus amplitude). Thus the model qualitatively captures this effect.

3.4 Processing of Flashed Stimuli. The second effect we consider is the presentation of briefly flashed high-contrast images. Figure 4a shows responses from neurons in area STSa to preferred and many nonpreferred

stimuli presented in randomly interleaved order for either 18 ms followed by a 93 ms gap, or for 102 ms followed by a 9 ms gap (data from Keysers et al., 2005). Yet despite the more than five-fold difference in stimulus duration, the neural responses are virtually identical (Keysers et al., 2005). A simple explanation for this observation would be a retinal afterimage. However, similar observations were reported when brief stimuli are immediately masked after their presentation (Rolls & Tovee, 1994), yielding retinal afterimages an unlikely explanation for the observed activity profiles.

Using the model of Figure 1, we compare the response to an 18 ms stimulus to the response to a 102 ms stimulus. In the model variant without recurrent and depressing synapses, the response duration and response amplitude clearly reflect the difference in stimulus duration (see Figure 4b). This is because the input is simply low-pass-filtered by the network. More precisely, in the limit where the filtering time constant is much shorter than the signal duration, the response duration reflects the stimulus duration, while in the limit where the filter time constant is much longer than the stimulus duration, the response amplitude reflects the stimulus duration.

In contrast, in the model with depressing recurrent connections, the model response is independent of stimulus duration, Figure 4c, as observed in the data. This is because the decay of the activity is dominated by the depression dynamics. With brief stimulus presentation, synapses are not yet depressed when the stimulus is removed, and hence the filtering time constant is still long. The response is sustained until the recurrent synapses are depressed out. For even shorter presentations (less than 10 ms), the response amplitude gradually decreases until the stimulus fails to propagate through the network and the stimulus presumably would not be perceived. The behavior in the population coding network is identical to this one-node-per-layer network (not shown).

The model's response to briefly flashed stimuli is shorter than seen in the neurophysiological data. The duration of the model's response is determined by the firing rates and how quickly the synapses depress, quantified by f in the model. We note again that we have deliberately made no attempt to fit model parameters to our data, preferring instead to take the parameters from the literature. While we could increase the value of f to match the duration of the response, this would simultaneously reduce the effect of the depression on response latency. Below we explore other possibilities.

Finally, we examined the network where the connections are not depressing (as in Figure 2d). Because of the high threshold in this network, the brief flash stimulus does not propagate, while the longer flash leads to a longer duration response (not shown). If the input is increased a factor three-fold, so that the brief stimulus does propagate, the half-width of the response to the brief stimulus is only 0.5 of the half-width of the response to the prolonged stimulus. Due to the high threshold, the response again strongly reflects stimulus duration.

3.5 Processing of Noisy Signals. The fact that the latency depends strongly on contrast shows that the time constant of the circuit adapts. This has advantages when processing noisy signals. To demonstrate the advantages of the adaptive network when processing noisy signals, we add gaussian white noise to step stimuli of low and high contrast (see Figure 5a, left and right, respectively). Without recurrence, the response is quick but also sensitive to noise, particularly evident to low-contrast stimuli (top traces). When nondepressing recurrence connections are included, the time constant of the circuit is slow. In this case, the noise is filtered out more, but the response at high-contrast is sluggish (middle traces). The circuit with depressing recurrent connections is both fast in response to high-contrast stimuli and filters the noise at low contrasts (bottom traces).

The processing of noisy signals is further quantified in Figure 5b. We measured the signal and its trial-to-trial variations 50 ms after response onset (when all networks have a strong response) and compared this to the absence of a stimulus. The resulting signal-to-noise ratio (SNR) normalized by stimulus amplitude (left) and the response latency (right) are plotted as a function of contrast. At low-stimulus contrasts, the SNR in the network with depressing recurrence (thick, solid curve) is superior to both the nondepressing recurrent network (dashed curve) and the nonrecurrent network (thin, solid curve). At high contrasts the nondepressing recurrent network has a higher SNR, but at the cost of increased latency. Note that the network with depressing recurrence at the highest contrasts will always have a longer latency than the nonrecurrent network because (1) synapses will require time to depress and (2) the recurrent synapses will never depress out completely.

The depressing recurrent network shows faster response latency than the nondepressing network across almost the entire contrast range, including at low contrasts where the normalized SNR is higher than that of the nondepressing network. This indicates the usefulness of adaptive networks. The initial network state has a long time constant, but as the responses develop, the network's time constant rapidly decreases, resulting in strong adaptive noise filtering with only a small cost in overall processing speed.

3.6 Rate Nonlinearity. In the model the firing rate is a smooth nonlinear function of the input current, Figure 6a (dashed curve). The nonlinearity (F/I curve) is expansive for small currents, modeling the effect of combining a noisy membrane potential with a firing threshold (Anderson, Lampl, Gillespie, & Ferster, 2000; Hansel & van Vreeswijk, 2002; Miller & Troyer, 2002). For large inputs, the nonlinearity is compressive, reflecting a maximal firing rate caused by the refractory period of the neurons.

While this nonlinearity is not essential to obtain contrast-dependent latency, the nonlinearity adds further realism to the model. First, the nonlinearity causes a realistic nonlinear relation between contrast and response, which becomes steeper the more layers are passed, Figure 6a. Responses in

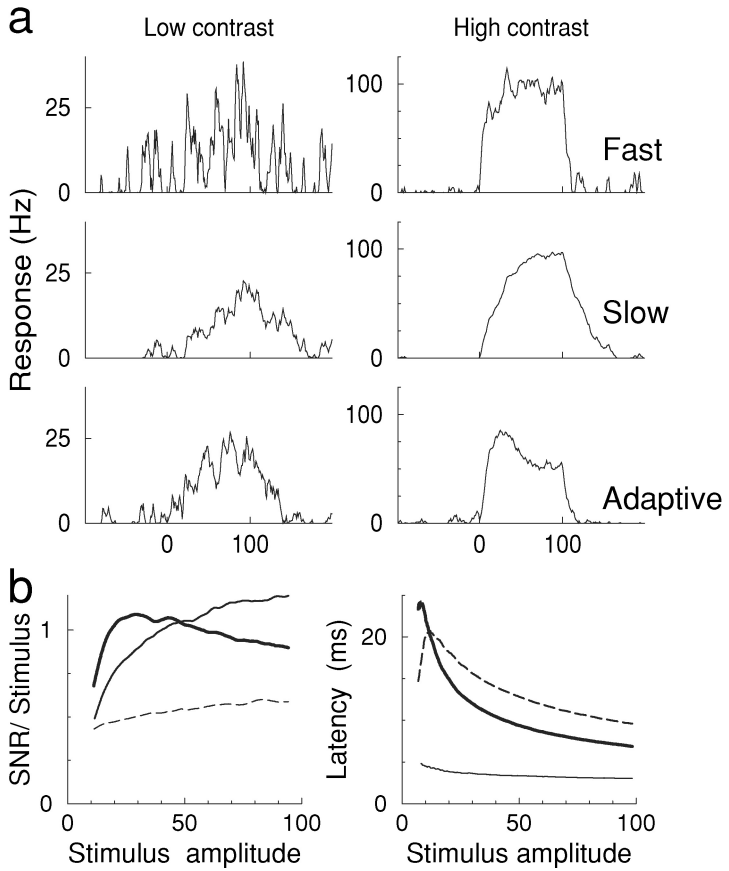

Figure 5: Adaptive noise filtering by the network. (a) The model's response in the first layer to a low-contrast (left) and high-contrast stimulus (right; note the difference in y-scale). The stimulus is a step stimulus (0..100 ms) to which gaussian white noise was added. A network without recurrent circuitry (top) reacts rapidly to signal transients but is noise sensitive. A recurrent network without depression filters out the noise but is sluggish at high contrasts (middle, recurrent feedback $g = 0.8$, depression factor $f = 1$). The network with recurrent depression (bottom) combines a rapid response at high contrast with a filtering of noise at low contrast. (b) The signal-to-noise ratio (left) and the latency to half-maximum (right) for the nonrecurrent network (thin, solid curve), nondepression recurrent network (dashed curve), and the network with depressing recurrence (thick curve). The signal-to-noise ratio was calculated across trials at 50 ms after stimulus onset with respect to the baseline response. For clarity, the signal-to-noise ratio was normalized by the stimulus amplitude.

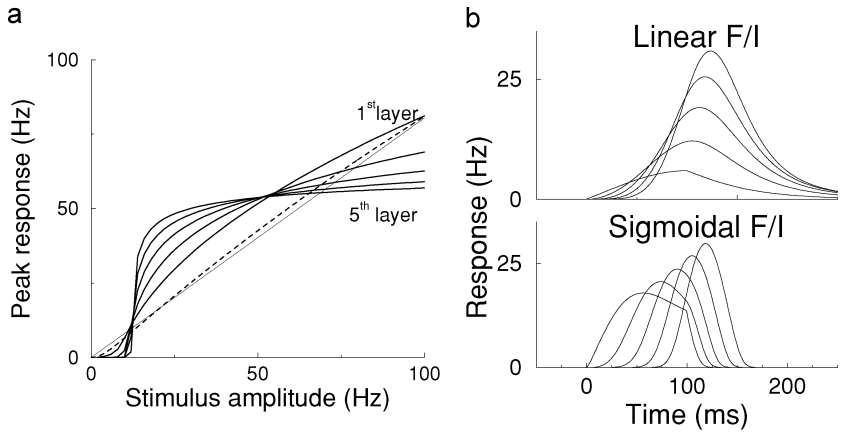

Figure 6: Normalization and nonlinearities in the model. (a) The maximum response amplitude as a function of the stimulus amplitude. Shown are the response in the first layer to the fifth layer of a five-layer network. In higher layers, the relation between stimulus and response is more strongly nonlinear. For comparison, the weakly nonlinear F/I curve of an individual node (dashed curve; see section 2) and a linear F/I curve (thin line) are also shown. (b) The effect of the nonlinearity on the firing rate. The firing rates of layers 1 to 5 are shown in response to a low-contrast step stimulus (100 ms duration). With a linear F/I curve, the response for low-contrast stimuli becomes temporally smeared (top). However, with a nonlinear F/I curve, the onset is less sluggish, and the response offset is particularly brisk (bottom), while the half-maximum onset latencies are similar to the linear model.

the higher layers are contrast independent. The recurrent connections amplify weak inputs, while high-contrast stimuli are amplified less (see also Figure 1c), enhancing the nonlinearity with every layer. Note that intermediate firing rates still occur when responses are part of a population code, but the response magnitude to a given stimulus becomes less dependent on stimulus contrast. Such normalizing behavior has been observed in subsequent stages of visual processing, where the contrast response function is almost linear in LGN but gradually steeper in V1 and MT (Sclar et al., 1990). This is a common feature of any layered model with a sigmoidal F/I curve.

It could be argued that the curve in higher layers is unphysiologically steep. However, real neurons will show heterogeneity in properties such as threshold and F/I curves, which will soften this steepness. Indeed, the steepness of the model's contrast-response relation (see Figure 6a) can be reduced by replacing the single nodes with a population of nodes with heterogeneity in the F/I curves and the connectivity (not shown).

The second effect of the nonlinearity is a steeper onset of the response and, in particular, a more rapid offset of the response as seen in the neurophysiological data. In Figure 6b, the activity in layer 1 to layer 5 is shown using a linear (top) and the nonlinear F/I curve (bottom). When the F/I curve is linear, the responses are temporally smeared at low contrast. In the nonlinear case, the half-maximum latencies are comparable, but the onset and the offset of the response are brisker. The mechanism behind the steeper onset resembles the spike-generation mechanism. At first, the recurrent feedback is hardly active as the input current does not lead to substantial activity. As activity builds up, the feedback gets disproportionately stronger, at which point the activity rapidly increases. The fast offset is observed because at the offset, the recurrent synapses throughout the network will be depressed; therefore, the network has a fast time constant, largely independent of the contrast.

3.7 NMDA and Sustained Responses. As mentioned above, the response in the higher layers is transient, even when the stimulus is maintained. One could argue that this in conflict with data, which often display sustained responses. We note that unlike the onset latency, the amount and time course of the sustained response vary greatly among cells. Furthermore, the late part of the response is often modulated by attention or higher area feedback (Roelfsema, Khayat, & Spekreijse, 2003). Finally, input might be coming from slower parallel inputs. From this point of view, a full model for sustained part might well be quite complicated.

Nevertheless, a more sustained response can be explained by including an NMDA type current in the model. To show this we added a current similar to equation 2.1 with a single exponential decay of 150 ms to all recurrent and feedforward synapses in the model. This current had 40% of the peak amplitude of the fast AMPA-like current. Voltage dependence of the NMDA conductance was not taken into account, as this requires a more explicit neuron model; results of such a network using integrate-and-fire neurons will be reported elsewhere.

When this NMDA-like current is included in the model, the response of the model is more sustained, while the other properties of the model remain largely intact. In Figure 7, the response in the fifth layer is analyzed, each time compared to the original model, without NMDA (thin lines). In Figure 7a the response to a 330 ms stimulus is plotted; the inclusion of the slow current lengthens the response, but the onset is unchanged. This is further illustrated in Figure 7b, where the onset and offset latency are plotted versus input amplitude (contrast). The onset latency is slightly delayed by the NMDA current, while the (half-maximum) offset latency is clearly much longer. The small effect on the onset latency is likely due to the response peaking later than without NMDA. Finally, the invariance of the response with regard to stimulus duration also occurs with the NMDA

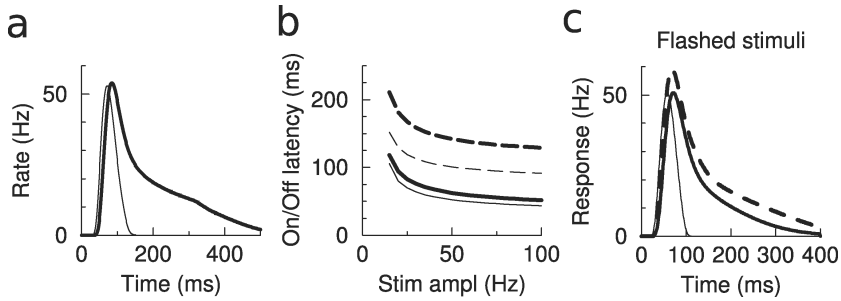

Figure 7: The effect of a slow NMDA-like current on the model. In each panel, the network with NMDA is shown in thick curves; the model without NMDA is shown in thin curves for comparison. (a) The response in the fifth layer to a 300 ms step stimulus. With NMDA, a sustained response results similar to what is seen in some recorded cells. (b) The onset latency (solid curve) and offset latency (dashed curve) as a function of stimulus. The onset latency is only slightly changed with NMDA, but the offset latency is extended considerably. (c) The response in the fifth layer to flashed stimuli as used in Figure 4. The NMDA component again lengthens the response, but the response remains insensitive to precise stimulus duration. Solid curve: response to brief stimuli; dashed curve: response to long stimuli.

current present, Figure 7c. In conclusion, including a slow conductance can explain the sustained response seen in the data.

4 Discussion

Cortical networks have abundant recurrent connections, the role of which has been speculated on in many models (Douglas et al., 1995; Ben-Yishai, Bar-Or, & Sompolinsky, 1995). Here we have shown that the short-term synaptic depression of the recurrent connections leads to an adaptive temporal integrator circuit. The model replicates a number of experimental findings: (1) contrast strongly affects the response latency, while the latency is only weakly coupled to response amplitude, (2) responses in higher areas are independent of stimulus duration for briefly flashed stimuli, and (3) the onset and the offset of the responses are brisk across stimulus contrasts. The network furthermore normalizes both the amplitude and temporal profile of the response (both become independent of the input contrast and duration), which is likely advantageous for subsequent processing. However, unlike gain control models that simply amplify weaker signals, it does so by dynamically adjusting the network time constant such that weaker signals are integrated over longer periods, thereby improving the resultant signal-to-noise ratio without sacrificing response time when the signal is strong.

The model is by no means meant to be a full model of the visual cortex in that it only tries to capture the dynamics of signal propagation. It ignores many known features of cortical circuitry. First, the model does not perform any computation and has an accordingly simple connectivity. Nevertheless, similar networks have been used to perform computations (van Rossum et al., 2002; van Rossum & Renart, 2004; Vogels & Abbott, 2005). Second, like most visual processing models, the connectivity ignores possible feedback from higher to lower areas, which could substantially complicate matters. Third, it ignores heterogeneity among the neurons: at low-contrast, latencies of neurons in the same area diverge substantially (see the error bars of Figure 2a). Nevertheless, we believe that the properties emerging in this model will hold in more involved models. As an example, we have observed similar dynamical properties in integrate-and-fire networks. The effects inherent in our model do not contradict, and indeed may act in concert with, recruitment of different feedback loops with changing stimulus contrast suggested by other models (Schwabe, Obermayer, Angelucci, & Bressloff, 2006), to explain contrast-dependent contextual interactions.

Given the many uncertainties about the nervous system and the many nonlinearities, it is hard to rule out all alternative explanations for the described phenomena. We believe that the proposed model is parsimonious and is consistent with the known architecture and physiological data. Nevertheless, its ultimate verification can be explored only in experiments.

Some studies have addressed contrast-dependent latency changes using a model that low-pass filters the input followed by a threshold (Bugmann & Taylor, 1993; Bair et al., 2002). These models can explain increased latency at lower contrasts and a difference between onset and offset latency. Our model variant without recurrence and without depression is an example of such a model (the filter is the synaptic time course, and a smooth threshold results from the F/I curve). Indeed, latency does depend on contrast, albeit more weakly (see Figure 2c). However, such models have rather short latencies and do not have an invariant response to brief flashed stimuli (see Figure 4b). An alternative model to explain long latencies would be one with depression feedforward but nondepressing recurrent connections (see Figure 2d); however, such a model is inconsistent with LGN-V1 data (Boudreau & Ferster, 2005).

Other studies have focused on temporal phase shifts in primary visual cortex as a function of contrast. In particular, Carandini and Heeger (1994) have suggested that inhibitory shunting feedback shortens the membrane time constant at high contrast. This idea and the mechanisms proposed here are not mutually exclusive, and the shunting model might help in explaining the contrast-dependent latencies. Furthermore, the resulting effective equations are quite similar in both models. Yet there are important differences that render the shunting model an unlikely sole explanation of the long latencies observed in higher areas. In the shunting model, the latency at low contrast is given by the membrane time constant, which is shortened by

inhibitory shunting feedback at high contrast. This seems hard to reconcile with studies that the effective membrane time constant is very short in vivo (Destexhe & Paré, 1999). Moreover, the even shorter synaptic time constant rather than the membrane time constant determines the circuit's dynamics (Knight, 1972; Treves, 1993). Finally, physiologically observed inhibition does not seem to match the shunting model (Ahmed, Allison, Douglas, & Martin, 1997; Anderson, Carandini, & Ferster, 2000). The mechanism of the model presented here is very different: first, there is no inhibition in the model; second, the time constant of the individual nodes is fixed and short. The long time constant is the result of the recurrence. As such there is in principle no upper bound to the latency. If it were possible to abolish both excitatory and inhibitory recurrent interactions, for example, by cooling, it would be possible to decide between the two models: a shunting model would predict a long time constant, while our model would predict a short time constant.

More recent models have examined V1 phase shifts using synaptic depression in either feedforward connections (Chance et al., 1998; Carandini et al., 2002), or feedforward and recurrent connections (Kayser et al., 2001). The extent of efficacy changes in the LGN-V1 pathway might, however, be limited, as the synapses are in a permanently depressed state due to the high background activity in LGN (Boudreau & Ferster, 2005) (while the same study shows that polysynaptic connections via other V1 cells do strongly depress). Interestingly, slower components in synaptic depression can be used to explain contrast adaptation on longer timescales (Chance et al., 1998), emphasizing the importance of synaptic dynamics on adaptive processing.

Appendix

Although analytical treatment of the coupled differential equations for firing rate and synaptic depression appears intractable, we here estimate the contrast dependence of the latency under simplifying assumptions. One major complication for solving the problem is that the release probability has a sigmoidal profile in time: when the input comes on, the release probability initially decays slowly as the activity is still low; next, the decay accelerates as the firing rate increases, while at later times, the firing rate decreases and the release probability reaches a steady state balanced by the recovery term.

To approximate these dynamics, we simply assume that the release probability P_{rel} decreases linearly in time when the node is activated $P_{rel}(t) = P_0[1 - 10^{-3} r_{\max}(1 - f)t]$, where $r_{\max}$ is the peak firing rate of the node for the given stimulus contrast (the factor 10^{-3} converts the firing rate from Hz into ms^{-1}). Furthermore, we assume that the F/I curve of the node is linear and limit ourselves to the first node. (The behavior for the deeper nodes is more complicated because of the normalization properties of the

network; see Figure 6). Stimulated with a step current I_{input}, the node's firing rate thus obeys according to equation 2.1, assuming $g = 1$, $P_0 = 1$,

$$\tau \frac{dr(t)}{dt} = -r(t) + I_{input} + r(t)(1 - kt/\tau)$$

where $k = 10^{-3}\tau r_{\max}(1 - f)$. Under the initial condition $r(0) = 0$, it has the solution

$$r(t) = I_{input}\sqrt{\frac{\pi}{2k}}\, e^{-\frac{kt^2}{2\tau^2}}\, i\,\mathrm{erf}\left(i\frac{t}{\tau}\sqrt{\frac{k}{2}}\right).$$

This function describes a transient pulse that rises quickly and decays slowly. We numerically extracted the half-maximum latency of $r(t)$ to be $t_{1/2} = 0.404\frac{\tau}{\sqrt{k}}$; hence, for the used parameters, $t_{1/2} \approx \frac{64}{\sqrt{r_{\max}}}$. A power law fit to the latency in the first layer in Figure 2b (plotted against the maximum firing rate at given contrast) yields $t_{1/2} = 90.r_{\max}^{-0.58}$, which is in reasonable agreement with the theory given the crudeness of the approximation.

Acknowledgments

We thank Guido Bugmann for discussion. M.vR. was partly supported by the EPSRC COLAMN Grant. M.vdM. was supported by the EPSRC and MRC through the Doctoral Training Centre in Neuroinformatics. D.X. was partly supported by EU framework grant (FP5-MIRROR) awarded to M.W.O. and D. I. Perrett.

References

Abbott, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science*, *275*, 220–224.

Ahmed, B., Allison, J., Douglas, R., & Martin, K. A. (1997). An intracellular study of the contrast-dependence of neuronal activity in cat visual cortex. *Cer. Cortex*, *7*, 559–570.

Albrecht, D. G. (1995). Visual cortex neurons in monkey and cat: Effect of contrast on the spatial and temporal phase transfer functions. *Vis. Neurosci.*, *12*, 1191–1210.

Albrecht, D. G., Geisler, W. S., Frazer, R. A., & Crane, A. M. (2002). Visual cortex neurons of monkeys and cats: Temporal dynamics of the contrast response function. *J. Neurophysiol.*, *88*, 888–913.

Albrecht, D., & Hamilton, D. (1982). Striate cortex of monkey and cat: Contrast response function. *J. Neurophysiol.*, *48*, 217–237.

Anderson, J. S., Carandini, M., & Ferster, D. (2000). Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *J. Neurophysiol.*, *84*, 909–926.

Anderson, J. S., Lampl, I., Gillespie, D. C., & Ferster, D. (2000). The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. *Science*, *290*, 1968–1972.

Bair, W., Cavanaugh, J. R., Smith, M. A., & Movshon, J. A. (2002). The timing of response onset and offset in macaque visual neurons. *J. Neurosci.*, *22*, 3189–3205.

Ben-Yishai, R., Bar-Or, R. L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci.*, *92*, 3844–3848.

Boudreau, C. E., & Ferster, D. (2005). Short-term depression in the thalamocortical synapses of cat primary visual cortex. *J. Neurosci.*, *25*, 7179–7190.

Bringuier, V., Chavane, F., Glaeser, L., & Frégnac, Y. (1999). Horizontal propagation of visual activity in the synaptic integration field of area 17 neurons. *Science*, *283*, 695–699.

Bugmann, G., & Taylor, J. G. (1993). A model for latencies in the visual system. In S. Gielen, & B. Kappen (Eds.), *Proc. ICANN'93* (pp. 165–168). Berlin: Springer.

Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, *264*, 1333–1336.

Carandini, M., Heeger, D. J., & Senn, W. (2002). A synaptic explanation of suppression in visual cortex. *J. Neurosci.*, *22*, 10053–10065.

Chance, F. S., Nelson, S. B., & Abbott, L. F. (1998). Synaptic depression and the temporal response characteristics of V1 cells. *J. Neurosci.*, *18*, 4785–4799.

Dayan, P., & Abbott, L. F. (2002). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Dean, A. F., & Tolhurst, D. J. (1986). Factors influencing the temporal phase of response to bar and grating stimuli for simple cells in the cat striate cortex. *Exp. Brain Res.*, *62*(1), 143–151.

Destexhe, A., & Paré, D. (1999). Impact of network activity on the integrative properties of neocortical pyramidal neurons. *J. Neurophysiol.*, *81*, 1531–1547.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. C., & Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*, 981–985.

Gautrais, J., & Thorpe, S. (1998). Rate coding versus temporal order coding: A theoretical approach. *Biosystems*, *48*, 57–65.

Gawne, T. J., Kjaer, T. W., & Richmond, B. J. (1996). Latency: Another potential code for feature binding in the striate cortex. *J. Neurophysiol.*, *76*, 1356–1360.

Gerstner, W. (2000). Population dynamics of spiking neurons: Fast transients, asynchronous state, and locking. *Neural Comp.*, *12*, 43–89.

Hansel, D., & van Vreeswijk, C. (2002). How noise contributes to contrast invariance of orientation tuning in cat visual cortex. *J. Neurosci.*, *22*, 5118–5128.

Hirsch, J. A., & Gilbert, C. D. (1996). Synaptic physiology of horizontal connections in the cat's visual cortex. *J. Neurosci.*, *11*, 1800–1809.

Kayser, A., Priebe, N. J., & Miller, K. D. (2001). Contrast-dependent nonlinearities arise locally in a model of contrast-invariant orientation tuning. *J. Neurophysiol.*, *85*, 2130–2149.

Keysers, C., Xiao, D.-K., Földiák, P., & Perrett, D. I. (2005). Out of sight but not out of mind: The neurophysiology of iconic memory in the superior temporal sulcus. *Cogn. Neuropsych.*, *22*, 316–332.

Knight, B. W. (1972). Dynamics of encoding in a population of neurons. *J. Gen. Physiol.*, *59*, 734–766.

Lee, B. B., Elepfandt, A., & Virsu, V. (1981). Phase of responses to moving sinusoidal gratings in cells of cat retina and lateral geniculate nucleus. *J. Neurophysiol.*, *45*(5), 807–817.

Loebel, A., & Tsodyks, M. (2002). Computation by ensemble synchronization in recurrent networks with synaptic depression. *J. Comput. Neurosc.*, *13*, 111–124.

Markram, H., & Tsodyks, M. (1996). Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature*, *382*, 807–810.

Maunsell, J. H., & Gibson, J. R. (1992). Visual response latencies in striate cortex of the macaque monkey. *J. Neurophysiol.*, *68*(4), 1332–1344.

Miller, K. D., & Troyer, T. W. (2002). Neural noise can explain expansive, power-law nonlinearities in neural response functions. *J. Neurophysiol.*, *87*, 653–659.

Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *J. Neurophysiol.*, *68*, 70–84.

Oram, M. W., Wiener, M. C., Lestienne, R., & Richmond, B. J. (1999). Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *J. Neurophysiol.*, *81*, 3021–3033.

Oram, M. W., Xiao, D., Dritschel, B., & Payne, K. R. (2002). The temporal resolution of neural codes: Does response latency have a unique role? *Phil. Trans. R. Soc. B*, *357*, 987–1001.

Raiguel, S. E., Xiao, D.-K., Marcar, V. L., & Orban, G. A. (1999). Response latency of macaque area MT/V1 neurons and its relationship to stimulus parameters. *J. Neurophysiol.*, *82*, 1944–1956.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neuro.*, *2*, 1019–1025.

Roelfsema, P. R., Khayat, P. S., & Spekreijse, H. (2003). Subtask sequencing in the primary visual cortex. *Proc. Natl. Acad. Sci.*, *100*, 5467–5472.

Rolls, E. T., & Tovee, M. J. (1994). Processing speed in the cerebral cortex, and the neurophysiology of visual masking. *Proc. R. Soc. B*, *257*, 9–15.

Saul, A. B. (1995). Adaptation aftereffects in single neurons of cat visual cortex: Response timing is retarded by adapting. *Vis. Neurosci.*, *12*, 191–205.

Schmolesky, M. T., Wang, Y., Hanes, D. P., Leutgeb, S., Schall, J. B., & Leventhal, A. G. (1998). Signal timing across the macaque visual system. *J. Neurophysiol.*, *79*, 3272–3278.

Schwabe, L., Obermayer, K., Angelucci, A., & Bressloff, P. C. (2006). The role of feedback in shaping the extra-classical receptive field of cortical neurons: A recurrent network model. *J. Neurosci.*, *26*(36), 9117–9129.

Sclar, G., Maunsell, J. H. R., & Lennie, P. (1990). Coding of image contrast in the central visual pathways of the macaque monkey. *Vis. Res.*, *30*, 1–10.

Shapley, R. M., & Victor, J. D. (1978). The effect of contrast on the transfer properties of cat retinal ganglion cells. *J. Physiol.*, *285*, 275–298.

Thomson, A. M., & West, D. C. (1993). Fluctuations in pyramid-pyramid excitatory postsynaptic potentials modified by presynaptic firing pattern and postsynaptic membrane potential using paired intracellular recordings in rat neocortex. *Neuroscience*, *54*, 329–346.

Treves, A. (1993). Mean-field analysis of neuronal spike dynamics. *Network*, *4*, 259–284.

Tsodyks, M. V., Pawelzik, K., & Markram, H. (1998). Neural networks with dynamic synapses. *Neural Comp.*, *10*, 821–835.

van Rossum, M. C. W., & Renart, A. (2004). Computation with populations codes in layered networks of integrate-and-fire neurons. *Neurocomputing*, *58*, 265–270.

van Rossum, M. C. W., Turrigiano, G. G., & Nelson, S. B. (2002). Fast propagation of firing rates through layered networks of noisy neurons. *J. Neurosci.*, *22*, 1956–1966.

Varela, J. A., Sen, K., Gibson, J., Fost, J., Abbott, L. F., & Nelson, S. (1997). A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. *J. Neurosci.*, *17*, 7926–7940.

Vogels, T. P., & Abbott, L. F. (2005). Signal propagation and logic gating in networks of integrate-and-fire neurons. *J. Neurosci.*, *25*, 10786–10786.

Wiener, M. C., Oram, M. W., Liu, Z., & Richmond, B. J. (2001). Consistency of encoding in monkey visual cortex. *J. Neurosci.*, *21*, 8210–8221.

Received June 13, 2007; accepted November 11, 2007.