Adaptive integration in the visual cortex by depressing recurrent cortical circuits

Mark C. W. van Roosum$^{1,*}$, Matthijs A. A. van der Meer$^{2}$, Dengke Xiao$^{3}$, and Mike W. Oram$^{3}$

January 26, 2008

$^{1}$Institute for Adaptive and Neural Computation
$^{2}$Doctoral Training Centre Neuroinformatics
School of Informatics, University of Edinburgh
5 Forrest Hill, Edinburgh, EH1 2QL, UK

$^{3}$School of Psychology, St. Andrews University
South Street, St. Andrews, Fife, KY16 9JU, UK

*Corresponding author. e-mail: mvanross@inf.ed.ac.uk

IN PRESS: Neural Computation

Abstract

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 100ms 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.
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, e.g. [Riesenhuber and Poggio, 1999]. However, stimulus contrast and stimulus duration typically also vary. Invariance to these changes cannot be obtained by pooling neurons, but 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 happens already in retina [Shapley and Victor, 1978], but also occurs in LGN [Lee et al., 1981], V1 [Albrecht and Hamilton, 1982, Dean and Tolhurst, 1986, Carandini and Heeger, 1994, Albrecht, 1995, Saul, 1995, Gawne et al., 1996, Bair et al., 2002, Albrecht et al., 2002], area MT [Raiguel et al., 1999], and in the anterior superior temporal sulcus (STS) [Oram et al., 2002]. The impact of contrast on response latencies becomes progressively larger in higher areas: as contrast is lowered, the latency of V1 responses increase from about 40ms to 75ms [Gawne et al., 1996], but in area STSa response latency increases from about 90ms to 250ms [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 and Taylor, 1993, Carandini and Heeger, 1994, Bair et al., 2002]. A number of previous models have included synaptic depression to explain temporal V1 response properties [Chance et al., 1998, Carandini et al., 2002, Kayser et al., 2001] (see also [Loebel and Tsodyks, 2002]), but typically have feed-forward connectivity only. This is problematic in light of evidence that synaptic depression in the drive from LGN to V1 is limited, while poly-synaptic connections via other V1 cells depress strongly [Boudreau and Ferster, 2005]. Furthermore, the large latency changes in higher areas are inconsistent with pure feed-forward models of visual processing. In models of spiking feed-forward 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 et al., 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 depends only weakly on the precise stimulus duration [Rolls and Tovee, 1994, Keysers et al., 2005]. This is also incompatible with linear feed-forward 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 et al., 1995], and short-term synaptic depression of these connections [Thomson and West, 1993, Markram and Tsodyks, 1996, Varela et al., 1997]. The abundant recurrent connections thus normalize signals by adaptively adjusting the network’s gain and time constant.

Methods

Model definition

We simulate a network model of nodes which are characterized by their firing rates, schematically depicted in Fig. 1c. Each node can be thought to represent the average firing rate of a small population of neurons, such as a micro-column. 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 and Abbott, 2002] for a discussion of this approximation). The net current \(I(t)\) received by a node is described with

\[
\tau \frac{dI(t)}{dt} = -I(t) + I_{\text{input}}(t) + gPr(t)r(t)
\]

(1)
where the time-constant $\tau$ (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 finally, the rightmost term is the recurrent feedback, indicated by the loops in Fig. 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 et al., 1998].

$$\tau_{depr} \frac{dP_{rel}(t)}{dt} = P_0 - [1 + \tau_{depr}r(t)(1-f)]P_{rel}(t) \quad (2)$$

where the parameter $\tau_{depr} = 500ms$ describes how quickly the synapse recovers towards its default release probability, given by $P_0 = 1$. The depression factor $f = 0.8$ describes how much each spike depresses the release probability ($P_{rel} \rightarrow fP_{rel}$). These parameter values were taken from the literature as they give an accurate description of depression in visual cortical slices [Abbott et al., 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

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

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

Feed-forward connections

Higher layers in the network receive their input from the previous layer. Like the recurrent connections, the synapses of these feed-forward connections are also subject to synaptic depression, however, the influence of this depression on the latency is weak (see Results). The input current to layer $i + 1$ in Eq. (1) is given by $I_{input}^{i+1}(t) = g_{ff}P_{rel}^{i+1}(t)r^{i}(t)$, where $P_{rel}^{i+1}(t)$ and $r^{i}(t)$ are release probability and the firing rate of the previous layer. An identical feed-forward gain $g_{ff}$ is used between all layers. This feed-forward gain is adjusted such that a sustained 50Hz stimulus in the input layer, evokes a response in the output layer with a peak rate of 50Hz. 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 feed-forward input.

Population coding network

For the study of population codes, we use a network with $N = 20$ neurons per layer with wrap-around 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"), 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 $w_{ik}^{ff} = gw_{ff} \cos(2\pi(i-k)/N)$, where $i$ and $k$ denote the lateral position of the nodes in the layers.

The feed-forward gain $g_{ff}$ is again determined by calibrating the peak responses.

The lateral connectivity matrix equals $w_{ik}^{lat} = (1 - \delta_{ij}) \cos(2\pi(i-j)/N)$. 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

$$I(t) = \sum_{j \neq i} w_{ij}^{lat} \cos(2\pi(j-i)/N)$$

$$+ \sum_{j} w_{ij}^{ff} \cos(2\pi(j-k)/N)$$

$$I_{stim,i} = A[\cos(2\pi(i-k)/N)]_+$$

$$I_{input,i} = g_{ff}P_{rel}^{i+1}(t)r^{i}(t)$$

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

$$\tau_{depr} \frac{dP_{rel}(t)}{dt} = P_0 - [1 + \tau_{depr}r(t)(1-f)]P_{rel}(t) \quad (2)$$

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

$$I_{stim,i} = A[\cos(2\pi(i-k)/N)]_+$$

$$I_{input,i} = g_{ff}P_{rel}^{i+1}(t)r^{i}(t)$$

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

$$\tau_{depr} \frac{dP_{rel}(t)}{dt} = P_0 - [1 + \tau_{depr}r(t)(1-f)]P_{rel}(t) \quad (2)$$

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

$$I_{stim,i} = A[\cos(2\pi(i-k)/N)]_+$$

$$I_{input,i} = g_{ff}P_{rel}^{i+1}(t)r^{i}(t)$$
\[ I_{input}(t) = \sum_j w^j_{ij} P^j_{rel}(t)r^j(t) + \sum_k w^k_{ik} P^k_{rel}(t)r^k(t), \]
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 non-depressing.

Axonal and dendritic propagation delays are not included in the model. Feed-forward 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 and Gilbert, 1990, Bringui er et al., 1999]. Including a fixed 10ms 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.

**Neurophysiological methods**

The experimental protocols have been described before [Oram et al., 2002]. Briefly, extra-cellular 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 300ms 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 110ms. Visual inspection of on-line rasters and the post-stimulus time histograms (PSTH) of each stimulus was used to select effective (preferred) and non-effective (non-preferred) stimuli.

To measure the effect of contrast on the response (Figs. 1, 2), gray-scale versions of preferred and non-preferred stimuli were presented for 333ms followed by an 333ms inter-stimulus interval. The different stimuli and different contrast levels were present in random order. The 100% Michelson contrast \((L_{\text{max}} - L_{\text{min}})/(L_{\text{max}} + L_{\text{min}})\) 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 \((75, 50, 25, 12.5, \text{and } 6.25\%)\) 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.

**Data analysis**  
Spike density functions were computed by smoothing a 1ms bin-width peri-stimulus time histogram with a Gaussian filter \(\text{sd} = 10\text{ms}\) for each stimulus at each contrast. Response magnitude was taken as the average firing rate in the 333ms 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 200ms prior to stimulus onset to 0 and the peak of the spike density function to 1, average across neurons, and re-normalizing to the range 0, 1 [Oram and Perrett, 1992].

The latency was extracted as the point at which the activity exceeded the baseline activity (estimated from 200ms before stimulus onset) by 3 standard deviations for at least 20ms. The latency was only accepted if the activity of the neuron in the 100ms 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 use 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 Results).
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 model and 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.

Contrast dependent latencies

We first study how response latency in higher visual areas depends on contrast. In Fig. 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 90ms 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 300ms (see also [Oram et al., 2002]). The response of a single neuron in area STSa of four different contrasts is shown in Fig. 1b. For this neuron, the response latency at 100% contrast was 80ms, which increased to 190ms 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 rise 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, while 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 Fig. 1c. At first we will 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 and Renart, 2004, Vogels and Abbott, 2005]. Based on the known anatomy and physiology each node receives both feed-forward 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 and connections in the model are subject to short-term synaptic depression, as observed in cortex [Thomson and West, 1993, Markram and Tsodyks, 1996, Varela et al., 1997]. Short term synaptic depression means that the synaptic response becomes weaker upon repeated stimulation, while after a period of rest the original strength is restored. We took the parameters of the synaptic depression from the literature (Methods), 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. Fig. 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 Fig. 1e we show the model response in layer 10 for four contrast levels. The latency increases substantially at low contrasts. Note that some 100ms 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 60ms, 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, Fig. 2a. For comparison we also show response latencies recorded in area V1 (data
Figure 1: Neural responses in higher visual areas to stimuli presented at different contrasts: data and model. a) Average normalized responses of 47 STSα 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 STSα neuron to a preferred stimulus (a “bio-hazard” sign) at different contrasts. Rastergrams and corresponding spike density functions (sd=10ms) 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 200ms 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 both from the previous layer and from recurrent excitation. Both the feed-forward 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.

from [Wiener et al., 2001, Oram et al., 1999]). The average response latency in area STSα increases by 33±3ms for each halving of stimulus contrast, which is significantly greater than in V1 where this was 8±0.8ms (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 Fig. 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} \left( \frac{r/r_{max}}{1-r/r_{max}} \right)^{1/n}, with parameters c_{50} = 0.5, n = 1.6,
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 6 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 layer (top curve). For the weakest stimuli, no latency is plotted in the deeper layer, because the stimulus fails to propagate deeply into the network.

c) Model latencies in layer 10 vs. 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 feed-forward 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 feed-forward connections are not depressing (thin dashed curve). A linear network without depression has a constant latency (straight line).

d) Model latencies in layer 10 vs. stimulus amplitude for a model in which the recurrent connections are not depressing. If the feed-forward connections are depressing, long contrast dependent latencies results (dashed curve). Even longer latencies result when the feed-forward connections are not depressing either (solid curve).

corresponding to LGN inputs [Sclar et al., 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 and Gibson, 1992, Schmolesky et al., 1998]. It has been estimated that between retina and AIT/STS, at least ten synapses must be traversed [Gautrais and Thorpe, 1998, Oram and Perrett, 1992]. With this in mind the latency in the model in layer 10 is comparable to the latency in area STS 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 non-linearity in the model (below).

Contributions to latency changes

Next, we explored the different contributions to the latency in the model. In Fig. 2c layer 10 latency is plotted versus contrast for a range of model conditions. In a feed-forward network, the non-linearity of the F/I curve by itself leads to a weak contrast dependent latency (thin solid curve). In this case each node filters and (smoothly) thresholds the signal; such models have been using to explain V1 latencies [Bair et al., 2002]. Adding depression in the feed-forward 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 feed-forward connections has again a small effect on the latency (thick dashed curve).

When the F/I curve is linear (i.e. $r = I$) and the synapses are non-depressing, the behavior of the network can easily be studied analytically. In that case, if recurrent feedback is absent, the latency would be proportional to $\tau$ times the number of layers crossed, where $\tau$ is the synaptic time constant, Fig. 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

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

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, Eqs 1 and 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, Fig. 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. Like above the non-linearity of the F/I curve and the feed-forward depression still lead to contrast dependent latencies, although the ratio of maximal and minimal latency is again smaller, Fig. 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 suggests the opposite [Boudreau and Ferster, 2005]. If the feed-forward connections are also not depressing, very long contrast dependent latencies result, Fig. 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 too high activity levels, the feed forward drive $g_{ff}$ is set much lower than in the other model variants (see Methods 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 non-linearity of the F/I curve and feed-forward 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 down 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 feed-forward networks.

**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 and Heeger, 1994; Albrecht, 1995; Gawne et al., 1996]. Comparable to the observations in V1, also in STSa and IT a non-preferred high contrast stimulus yields a small response, but with a short latency. This is illustrated in Fig. 3a (see also [Oram et al., 2002; Oram and Ferrett, 1992]).

To examine the dependence of response latency on response magnitude and stimulus contrast in the model, we implement a population coding network, Fig. 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 (Methods). Subsequent layers are connected to each other with a weight matrix which 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, Fig. 3c: (left and right). In the network with depressing recurrence, the latency is again strongly contrast dependent, Fig. 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 (Fig. 3c, lower left) is higher than the activity of the edge nodes at high contrast (Fig. 3c, lower right), yet the latency at low contrast is about two-fold 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 non-linear, a small contrast dependent latency remains, Fig. 2c. In the population coding network, the latency is again weakly
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 rates. 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).

contrast dependent, Fig. 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, non-preferred 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 high activity. Secondly, 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 67±7% of the variability of response latency and only 33±3% of the variability in spike count. Conversely, stimulus identity accounted for 69±6% of the variability in spike count and only 20±5% of the variability on response latency. Thus, in areas STS 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.

Processing of flashed stimuli

The second effect we consider is the presentation of briefly flashed high contrast images. Fig. 4a shows responses from neurons in area STS to preferred and non-preferred 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 5-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 and Tovee, 1994], yielding retinal afterimages an unlikely explanation for the observed activity profiles.

Using the model of Fig. 1 we compare the response to a 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 (Fig. 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, Fig. 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 (< 10 ms), the response amplitude gradually decreases until the stimulus fails to propagate through the network and the stimulus would presumably 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

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 non-preferred 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.
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, 100ms) to which Gaussian white noise was added. A network without recurrent circuitry (bottom) 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 (top) 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 non-recurrent network (thin, solid curve), non-depression recurrent network (dashed curve), and the network with depressing recurrence (thick curve). The signal to noise ratio was calculated across trials at 50ms after stimulus onset, with respect to the baseline response. For clarity the signal-to-noise ratio was normalized by the stimulus amplitude.

attempt to fit model parameters to fit 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 Fig. 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 3-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.

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 (Fig. 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 non-depressing 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 Fig. 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 non-depressing recurrent network (dashed curve) and the non-recurrent network (thin, solid curve). At high contrasts the non-depressing 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 non-recurrent 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 non-depressing network across almost the entire contrast range, including at low contrasts where the normalized signal to noise ratio is higher than that of the non-depressing 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.

**Rate non-linearity**

In the model the firing rate is a smooth non-linear function of the input current, Fig. 6a (dashed curve). The non-linearity (F/I curve) is expansive for small currents, modeling the effect of combining a noisy membrane potential with a firing threshold [Anderson et al., 2000b, Hansel and van Vreeswijk, 2002, Miller and Troyer, 2002]. For large inputs the non-linearity is
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 300ms 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 Fig. 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.

This non-linearity, reflecting a maximal firing rate caused by the refractory period of the neurons.

While this non-linearity is not essential to obtain contrast dependent latency, the non-linearity adds further realism to the model: First, the non-linearity causes a realistic non-linear relation between contrast and response, which becomes steeper the more layers are passed, Fig. 6a. Responses in the higher layers are contrast independent. The recurrent connections amplify weak inputs, while high contrast stimuli are amplified less (see also Fig. 1c), enhancing the non-linearity 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 more steep 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 un-physiologically 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 (Fig. 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 non-linearity is a steeper onset of the response and, in particular, a more rapid offset of the response as seen in the neurophysiological data. In Fig. 6b the activity in layer 1 to layer 5 is shown using a linear (top) and the non-linear F/I curve (bottom). When the F/I curve is linear, the responses are temporally smeared at low contrast. In the non-linear 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.

**NMMA 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 varies greatly amongst cells. Furthermore, the late part of the response is often modulated by attention or higher feedback, e.g. [Roelfsema et al., 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 a NMDA type current.
in the model. To show this we added a current similar to Eq.(1) with a single exponential decay of 150ms to all recurrent and feed-forward 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 Fig. 7 the response in the 5th layer is analyzed, each time compared to the original model without NMDA (thin lines). In Fig. 7a the response to a 3300ms stimulus is plotted, the inclusion of the slow current lengthens the response, but the onset is unchanged. This is further illustrated in Fig. 7b, where the onset and offset latency are plotted vs. 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 w.r.t. stimulus duration also occurs with the NMDA current present, Fig. 7c. In conclusion, including a slow conductance can explain the sustained response seen in the data.

Discussion

Cortical networks have abundant recurrent connections, the role of which has been speculated about in many models e.g. [Douglas et al., 1995, Ben-Yishai et al., 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 and Renart, 2004, Vogels and Abbott, 2005]. Secondly, like most visual processing models, the connectivity ignores possible feedback from higher to lower areas, which could substantially complicate matters. Thirdly, it ignores heterogeneity amongst the neurons: at low contrast latencies of neurons in the same area diverge substantially (see error bars of Fig. 2a). Nevertheless, we believe that the properties emerging in this model will hold in more involved models. As an example, we have observed very 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, e.g. [Schwabe et al., 2006], to explain contrast dependent contextual interactions.

Given the many uncertainties about the nervous system and the many non-linearities, 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 only be explored in experiments.

Some studies have addressed contrast dependent latency changes using a model that low-pass filters the input followed by a threshold [Bugmann and 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 (Fig. 2c). However, such models have rather short latencies and do not have an invariant response to brief flashed stimuli (see Fig. 4b).

An alternative model to explain long latencies would be one with depression feed-forward but non-depressing recurrent connections (Fig 2d), however, such a model is inconsistent with LGN-V1
data [Boudreau and Ferster, 2005].

Other studies have focused on temporal phase shifts in primary visual cortex as a function of contrast. In particular, Carandini and Heeger have suggested that inhibitory shunting feedback shortens the membrane time constant at high contrast [Carandini and Heeger, 1994]. 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 which 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 and 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 et al., 1997, Anderson et al., 2000a]. The mechanism of the model presented here is very different: firstly, there is no inhibition in the model; secondly, 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, e.g. 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 with in either feed-forward connections [Chance et al., 1998, Carandini et al., 2002], or in feed-forward and recurrent connections [Kayser et al., 2004]. 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 and Ferster, 2005] (while the same study shows that poly-synaptic connections via other V1 cells do strongly depress). Interestingly, slower components in synaptic depression can be used to explain contrast adaptation on longer time scales [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, Fig. 6). Stimulated with a step current $I_{input}$, the node’s firing rate thus obeys according to Eq. (1); assuming $g = 1, P_0 = 1$

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

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

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

This function describes a transient pulse which rises quickly and decays slowly. We numerically extracted the half maximum latency of $r(t)$ to be $t_{1/2} = 0.405 \sqrt{\tau}$, hence for the used parameters $t_{1/2} \approx 64 \sqrt{\tau}$. A power-law fit to the latency in the first layer in Fig. 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. MvR was partly supported by the EPSRC COLUMIN Grant. MedM was supported by the EPSRC and MRC through the Doctoral Training Centre in Neuroinformatics. DX was partly supported by EU framework grant (FP5-MIRROR) awarded to MWO and D.I. Perrett.

References


