Symposium on Dynamics of Neural Processing, Washington DC, 6-8 June 1994.
Extended Abstract Book, 132-136

ROLE OF SHORT-TERM MEMORY IN NEURAL INFORMATION PROPAGATION

Guido Bugmann and John G. Taylor*
School of Computing, University of Plymouth, Plymouth PL4 8AA, United Kingdom
* Department of Mathematics, King's College London, London WC2R 2LS, United Kingdom
email: gbugmann@sc.plym.ac.uk, Fax: (+44) 752 23 25 40

Introduction: Short-term memory (STM) refers to the temporary retention of information needed for a later action. Short-term memory neurons exhibit a sustained firing during the retention time and are tuned to specific information. Initially extensively described in the prefrontal cortex (area 46) [1], STM-neurons have since been found in many areas of the brain; e.g. in the motor cortex (area 4) [2] which receives projections from area 46 via area 6, in area IT of the visual system which projects to area 46, and in area V1 which, via a certain number of relays, projects to IT [3]. These multiple findings suggest that STM may not be a localised special function but may be involved at all levels of neural information processing.

In this paper we argue that STM is needed for enabling neural information to propagate from layer to layer while preserving the level of firing frequency. This idea is demonstrated with a model of a feedforward pyramidal neural network where all neurons act as coincidence detectors and respond with a prolonged sustained firing to each coincidence of input spikes. The form of this neural network is sufficiently general to be compatible with most hypothesis on neural information processing and is valid for most information processing streams. It also allows exact mathematical analysis [Bugmann and Taylor, in preparation]. Some properties of the model such as the capability for parallel and serial pattern recognition, sequence recognition, timing and others are presented elsewhere [4]. In this paper we use the model to discuss two characteristics of visual information processing: latencies and masking phenomena.

The need for STM arises from two observations. Firstly, neurons in the visual system produce very irregular, near random, spike trains. This has been shown to be incompatible with models of the neuronal function based on temporal integration of EPSP's [5]: irregular spike trains can only be produced if the neuron fires in response to randomly occurring coincidences of inputs spikes. Secondly, the conditions for coincidence detection can be analysed using the Leaky Integrate-and-Fire (LIF) model of a neuron [6]. The number of neurons in a layer connected to a neuron in the next layer, e.g. from LGN to V1, is approximately 240 [7]. Even if only a fraction of these neurons are activated by a given stimulus, the input current in the target cell is relatively continuous and shows small fluctuations around its average value. The small current peaks above the average are due to the synchronous arrival of a more or less large number of input spikes. If a neuron has to operate as coincidence detector, only these peaks must be able to trigger an output spike. This implies that the average input current must lie below the current threshold [6,8]. Under these conditions, the firing frequency of the LIF is very small [6], there is a considerable input-output frequency drop and the neural information vanishes after propagating through a few layers. As the level of firing remains relatively constant throughout the visual system, there must exist some amplification system, which, as we will see below, has STM properties.

The simplest form of amplification is provided by a positive feedback loop [4]. It has been shown that, in the visual cortex, most cells are imbedded in a local network providing a very effective excitatory feedback which, if inhibition is inactivated, can cause a prolonged sustained firing [9]. The stand-alone firing frequency of the neuron (frequency of coincidences of feed-forward input spikes) determines which gain of the feedback loop is required to restore a standard output firing level. If the gain is very large, most spikes produced by the neuron are actually generated by the local network. Ending the input activity has only a minor effect and the neuron continues to fire [4]. This is a form of short-term memory where the retained information is the past occurrence of a given input condition. We may note that, for consistency, the question of the irregularity of the spike trains should now be re-discussed in term of local circuit effects. This is beyond the scope of this paper, however.

Model: After having discussed the need for STM and the existence of the necessary circuit, we analyse the information propagation in a very simple model implementing these ideas. In the example in figure 1, all neurons have *m* inputs from the preceding layer. Neurons operate in discrete time steps and fire only in response to the coincidence of all *m* input spikes. As soon as a coincidence is detected, the neuron starts firing a random spike train with a probability P_1 of generating a spike at each time step. This reproduces in a simplified way the effect of the local circuit. It results in a binary operation, the neuron being silent or firing with a probability P_1. Each spike carries therefore the same information and the target neuron performs a multiple AND-function. The layer 0 is the input layer representing, for instance, the retinal ganglion cells. These respond to a brief flash of light with a prolonged burst of spikes spike characterised by a latency and a jitter in starting time [10]. We reproduce the jitter in following way: at time 0, a neuron in layer 0 is set to fire with a probability $P_0 < P_1$. As soon as the first spike is produced, the neuron is switched to sustained firing with firing probability P_1, which reproduces the prolonged burst. The jitter in starting time, determined by P_0, can be observed in figure 2. The latency is not modelled. For consistency we have introduced a reset function which stops the sustained firing of all *m* neurons in a layer *n* when their target neuron in layer *n+1* has started firing. This enables neurons in layer *n* to respond to new information. It also reduces to a strict minimum the duration of the sustained firing, thus minimising energy consumption while ensuring information propagation. Within this scheme, sustained firing is hidden as most neurons produce only a relatively short burst of spikes in response to a stimulus. Figure 2 shows the stochastic nature of the minimum duration of firing needed to ensure information propagation. When the reset function disables permanently the neurons, post-stimulus histograms (PSTH) show a propagation of a firing probability wave with a temporal overlap of the activity over several layers, as in figure 2. If we allow the neurons to restart their sustained firing as soon as a new coincidence occurs, the simulations produce temporal profiles of PSTH's with complex oscillatory components very similar to those observed in real neurons (unpublished results).

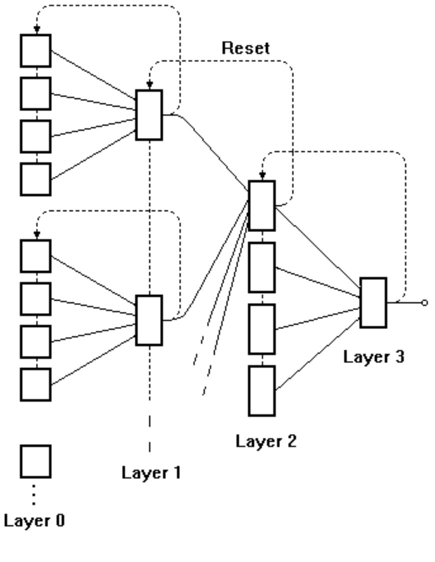

Figure 1. Pyramidal Neural network used in our simulations. All neurons have 4 inputs and, in response to a coincidence of 4 input spikes, initiate sustained firing. When firing, a neuron stops the sustained firing of its 4 input neurons in the previous layer.

Latencies: In the model, there is only one time-step propagation delay between layers. However, simulations show a much larger increase in latency as the information propagates from layer to layer. Latencies are the time of reaching $P_1/2$ in the histograms, in the case of no reset. Latencies are due to the fact that a neuron in layer *n* has no chance to start firing before all its *m* inputs in layer *n-1* are in a state of sustained firing. Even when all inputs are firing with a probability P_1, a target neuron must still wait for a coincidence to occur. Figure 3 shows the inter-layer latencies ΔLn as a function of P_1, and the curve predicted by the theoretical analysis of the model [11; Bugmann and Taylor, in preparation]. A notable effect is the saturation towards a minimum latency ΔLmin as P_1 increases. For large values of P_1, a coincidence occurs almost immediately after all inputs are firing. The main cause of the latencies is then the jitter in layer 0. Figure 4 shows how the minimum latency depends on P_0. This somehow surprising result states that whatever layer you are looking at, above a certain level of sustained firing, the time taken for the information to propagate from the previous layer depends mainly on the jitter in the very first layer of the pyramid. In the case of the visual system, it is therefore possible that the retinal jitter controls the information processing speed at all stages. Experimental manipulations of the jitter may lead to considerable effects. As the total latency in a given layer is the sum of all previous inter-layer latencies, a small change in the standard deviation of the retinal onset time is multiplied by the number of layers through which the information propagates.

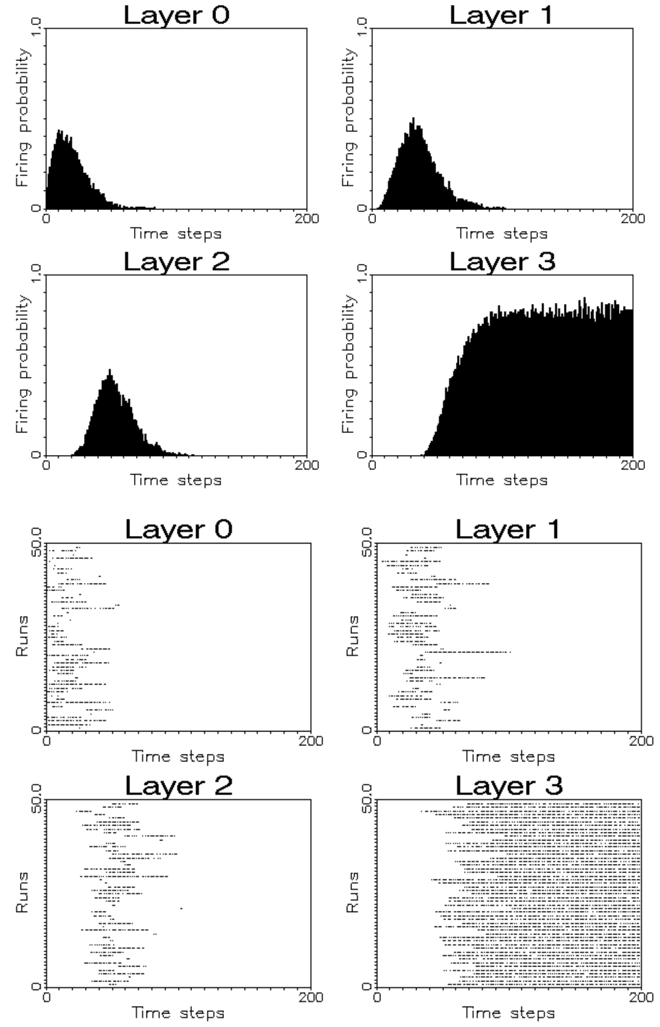

Figure 2. Results of simulations of the network in figure 1. Post stimulus histograms and spike rasters are shown for the first neuron in each layer. Each neuron is part of the 4 inputs of the neuron in the next layer. We used $P_0 = 0.08$ and $P_1 = 0.8$. The histograms are obtained from 200 runs.

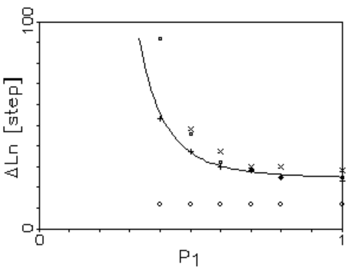

Figure 3. Interlayer latency difference ΔLn in dependence on P_1, using $P_0 = 0.4$. The symbols correspond to different layers. The bottom row of circles is the latency in the layer 0. The full line is the theoretical function
$\Delta Ln = A + B P_1^{-m}$

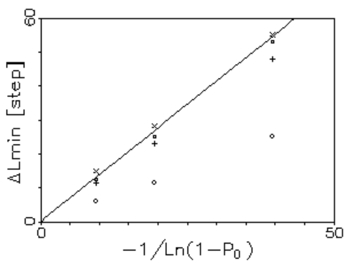

Figure 4. Minimum interlayer latency difference ΔLmin in dependence on P_0. The data measured by simulations using $P_1 = 1$ (symbols) and the theoretical function (full line) are represented as a function of $-1/\text{Log}(1-P_0)$. The theoretical slope is Log(*m*).

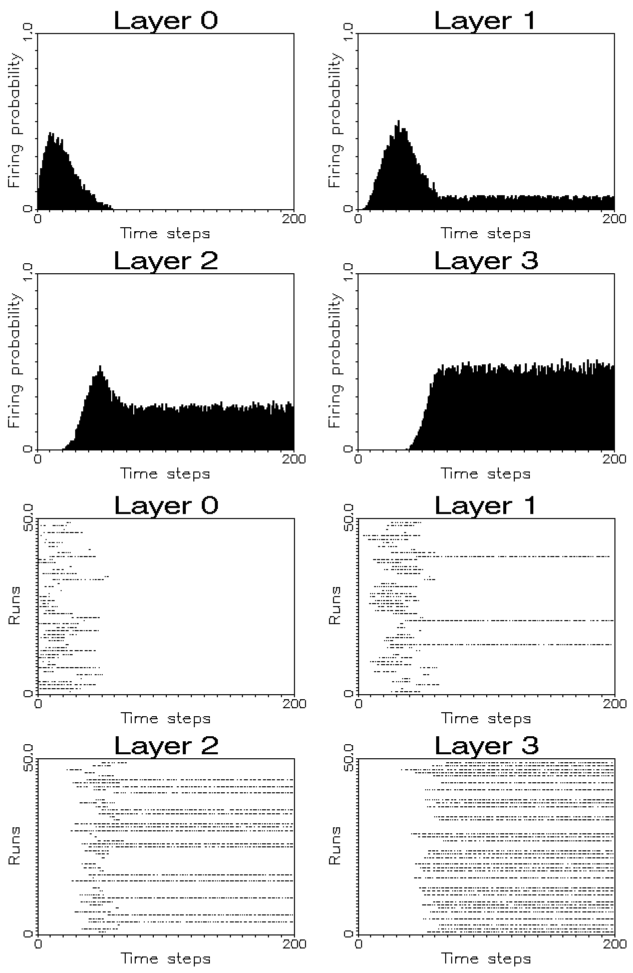

Figure 5. Post stimulus histograms and spike rasters observed when the sustained firing in layer 0 is interrupted at the time step 60. We used $P_0 = 0.08$ and $P_1 = 0.8$.

Backward masking. In backwards masking experiments, a subject has to detect, in a first image presented briefly, the presence or the absence of some visual pattern among an array of distractors, e.g. a horizontal bar among vertical bars. The first visual stimulus is shown at time 0 during typically 33ms, a blank screen follows and, at time SOA (Stimulus Onset Asynchrony), a second stimulus is displayed containing, for instance, an array of crosses (the mask) [12]. As the delay SOA becomes smaller, the performance of the subject decreases and reaches the chance level where he or she produces 50% of correct responses. Examples of psychometric response curves are shown on figure 6, rescaled between 0 and 1. The intriguing feature of backwards masking is that a later stimulus can affect the processing of earlier information, and it does so in a probabilistic way.

In our model, information remains localised in layer 0 during a certain time after the onset of the stimulus, until the target neuron in layer 1 has seen a coincidence and starts firing. During this time, it is possible for new information to perturbate older information. Based on electrophysiological recording, it has been suggested that the effect of the mask is to reduce the duration of the spike trains carrying sensory information [13]. We have implemented this idea by limiting the duration of the sustained firing in layer 0. Figure 5 shows that the ending of the firing in layer 0 at a predetermined time is sometimes premature. It results in a number of runs where neurons in layer 1 fail to start firing. In these cases, the neuron in layer 3 cannot fire and inhibit its other inputs in layer 2. This suggests that prolonged sustained firing may be the electrophysiological correlate of masking.

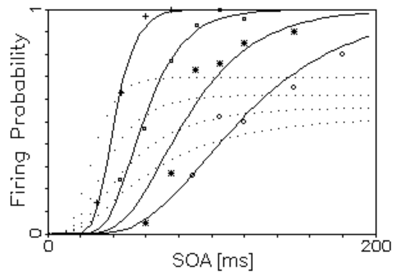

Figure 6. Psychometric curves produced by our model of backward masking. Symbols: Data from fig. 4 in Bergen and Julesz [14]. Full lines: probability for the neuron in layer 3 to start sustained firing in dependence on the stop time SOA in layer 0 (assuming 1 time step = 1 ms). Dotted lines: Time dependent firing probabilities of neurons in layer 1 which produce the best fit to the experimental curves. Other parameters: $P_0 = 0.084$, m=6, N=4.

To determine the psychometric curve predicted by our model, we have assumed that the response of the subject is based on information stored in short-term memory, for instance the neuron in layer 3. If this neuron fails to enter in sustained mode, which is equivalent to any of the N neurons in layer 1 failing to do so, then the subject has no information on the nature of the first image and must answer at random. We have calculated the theoretical probability of correct responses in dependence on the stop time SOA in layer 0 and the parameters of the model, P_0, P_1, the number *m* of inputs and the number *N* of neurons in layer 1 which must be activated. Figure 6 shows that the resulting function [Bugmann and Taylor, in preparation] produces a good fit to a set of experimental data using only P_1 as an adjustable parameter. In this model, the threshold value of SOA below which the subject has no information, is determined by the initial jitter, controlled by the parameter P_0. Other models do not predict the threshold [12].

Conclusion. We have explored the hypothesis that sustained firing results from the use of the local network for amplification. However, sustained firing may also be the expression of a more complex computational process, possibly involving lateral information propagation. The results on the role of the jitter suggest that synchronisation between input neurons is likely to increase the propagation velocity of the information carried by these neurons. The application of the model to masking deserves further thought, the temporal overlap of information contributing probably also to the temporal fusion threshold. Inhibitory feedback has been used in several models (see references in [11]). It affects neither the velocity of information propagation, nor masking effects and is not critical for these phenomena. On the other hand, a purely local control of the duration of the sustained firing would cause failures of information propagation, because some neurons in a layer *n* may stop firing before their target in layer *n+1* has had time to detect a coincidence. This is seen in the masking model. Therefore, the question of how the duration of the sustained firing is controlled needs to be addressed carefully.

References.

1. *Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex.* S.Funahashi, C.J. Bruce and P. Goldman-Rakic, J. Neurophys. 61, 331-349, 1989.

2. *Motor cortical activity in a memorised delay task.* N. Smyrnis, M. Taira, J. Ashe and A.P. Georgopoulos, Exp. Brain Res., 92, 139-151, 1992

3. *Inferotemporal units in selective visual attention and short-term memory.* J.M. Fuster, J. Neurophys., 64, 681-697, 1990.

4. *A stochastic short-term memory using a pRAM neuron and its potential applications.* G. Bugmann and J.G. Taylor, in Beale R. and Plumbley M.D. (eds) "Recent advances in neural networks", Proc. of BNNS'93, Prentice Hall, in press.

5. *The high irregular firing of cortical-cells is inconsistent with temporal integration of random EPSP's.* W.R. Softky and C. Koch, J. Neurosci., 13, 334-350, 1993.

6. *Summation and multiplication: Two distinct operation domains of Leaky Integrate-and-Fire neurons.* G. Bugmann, Network, 2, 489-509, 1991; *Multiplying with neurons: Compensation for irregular input spike trains by using time-dependent synaptic efficiencies.* G. Bugmann, Biol. Cybern., 68, 87-92, 1992.

7. *Polyneural innervation of spiny stellate neurons in cat visual cortex.* B. Ahmed, R.J. Douglas, K.A. Martin and C. Nelson, J. Comp. Neurol., 341, 39-49, 1994.

8. *The temporal noisy-leaky integrator neuron model*. C. Christodoulou, G. Bugmann, T. Clarkson and J.G. Taylor, in same book as [4], in press.

9. *A functional microcircuit for cat visual cortex.* R. Douglas and K.A. Martin, J. Physiol., 440, 735-769, 1991.

10. *Variation in the response latency of cat retinal ganglion cells.* W.R Levick, Vision Res., 13, 837-853, 1973.

11. *A model for latencies in the visual system.* G. Bugmann and J.G. Taylor, in Proc. 3rd Conf. on Artificial Neural Networks (ICANN'93, Amsterdam), Gielen S. and Kappen B. (eds), p.165-168, 1993.

12. *Time course of perceptual discrimination and single neuron reliability*. E. Zohary, P. Hillman and S. Hochstein, Biol. Cybern., 475-486, 1990.

13. *Visual masking: Mechanisms and theory*. G. Felsten and G.S. Wasserman, Psychological Bulletin, 88, 329-353, 1980.

14. *Rapid discrimination of visual patterns*. J.R. Bergen and B. Julesz, IEEE Trans. SMC-13, 857-863, 1983.