The Journal of Neuroscience, August 1993, 13(8): 3406–3420

# A Spiking Network Model of Short-Term Active Memory

**David Zipser,[1] Brandt Kehoe,[2] Gwen Littlewort,[1] and Joaquin Fuster[3]**

[1]Department of Cognitive Science, University of California at San Diego, La Jolla, California 92093, [2]Department of Physics, California State University, Fresno, California 93740, and [3]Brain Research Institute and Department of Psychiatry, School of Medicine, University of California, Los Angeles, California 90024

**Studies of cortical neurons in monkeys performing short-term memory tasks have shown that information about a stimulus can be maintained by persistent neuron firing for periods of many seconds after removal of the stimulus. The mechanism by which this sustained activity is initiated and maintained is unknown. In this article we present a spiking neural network model of short-term memory and use it to investigate the hypothesis that recurrent, or "re-entrant," networks with constant connection strengths are sufficient to store graded information temporarily. The synaptic weights that enable the network to mimic the input–output characteristics of an active memory module are computed using an optimization procedure for recurrent networks with nonspiking neurons. This network is then transformed into one with spiking neurons by interpreting the continuous output values of the nonspiking model neurons as spiking probabilities.**

**The behavior of the model neurons in this spiking network is compared with that of 179 single units previously recorded in monkey inferotemporal (IT) cortex during the performance of a short-term memory task. The spiking patterns of almost every model neuron are found to resemble closely those of IT neurons. About 40% of the IT neuron firing patterns are also found to be of the same types as those of model neurons.**

**A property of the spiking model is that the neurons cannot maintain precise graded activity levels indefinitely, but eventually relax to one of a few constant activities called fixed-point attractors. The noise introduced into the model by the randomness of spiking causes the network to jump between these attractors. This switching between attractor states generates spike trains with a characteristic statistical temporal structure. We found evidence for the same kind of structure in the spike trains from about half of the IT neurons in our test set. These results show that the behavior of many real cortical memory neurons is consistent with an active storage mechanism based on recurrent activity in networks with fixed synaptic strengths.**

***[Key words: short-term memory, neural network model, inferotemporal cortex, memory model, spiking model neurons, attractor dynamics]***

Animals have memories with retention times ranging from fractions of a second to a lifetime. Two different ways of maintaining information in memory have been proposed: one passive, or latent, and the other active. In passive storage, information about the item is maintained in modified values of physiological parameters such as synaptic strength. Neural activity is required only during loading and retrieval, but not for maintenance. In active storage, information is preserved by maintaining neural activity throughout the time it must be remembered. There is experimental evidence that both of these information storage strategies are used in higher animals. Latent information apparently can be stored in the brain indefinitely, but active information can be maintained only for relatively short times, perhaps a few tens of seconds; thus, it is reasonable to refer to this kind of memory as *active short-term memory.* Little is known about the mechanism of active information storage. In the research described here we have investigated a possible mechanism for information storage in active short-term memory by using a spiking neural network model whose behavior can be directly compared to experimental findings.

Lesion and brain cooling studies have identified several cortical areas that are required for short-term memory tasks, such as *delayed match to sample* or *delayed response,* but not required for versions of the same tasks without a delay. Areas devoted to specific modalities, such as the inferotemporal (IT) cortex for vision or posterior parietal for touch, are required only for tasks dealing with stimuli of those modalities. However, the prefrontal cortex appears to be required for all memory tasks that involve a delayed motor response (Bauer and Fuster, 1976; Fuster, 1985, 1989; Fuster et al., 1985; Goldman-Rakic, 1987; Quintana et al., 1989).

Recordings of single-unit activity in monkeys performing short-term memory tasks have been carried out for over two decades (Fuster and Alexander, 1971; Fuster, 1973; Niki, 1974; Fuster and Jervey, 1982; Fuster et al., 1982; Quintana et al., 1988; Koch and Fuster, 1989; Funahashi et al., 1990). This work has demonstrated that many neurons in the areas required for short-term memory are associated with the memory task in some way. Three main criteria have been used to show that neurons are memory relevant. These are (1) systematically altered activity during the delay period, while information about a stimulus

Received July 30, 1992; revised Feb. 9, 1993; accepted Feb. 10, 1993.

This work was supported by Office of Naval Research Grant 14-89-J-1805 to J.F., and by System Development Foundation Grant G359D and National Institute of Mental Health Grant MH45271 to D.Z. J.F. holds Research Scientist Award MH 25082. We thank Bill Bergerson and Bradford Lubell for preparing hardware and software for analysis of the experimental data; we also thank Mark Bodner for useful comments on the manuscript.

Correspondence should be addressed to David Zipser, Department of Cognitive Science, University of California, San Diego, 9500 Gilman Drive, Department 0515, La Jolla, CA 92093-0515.

must be retained in the absence of the stimulus, (2) failure of memory relevant neurons to respond to stimuli that need not be memorized, generally because of absence of reward expectation (Fuster, 1973, 1990), and (3) correlation between error on a memory task and response failure of memory-relevant neurons (Fuster, 1973).

An example of sustained activity during delay periods of various lengths in a delayed saccade task is shown in Figure 1*A*. The direction and magnitude of the saccade are the same for all trials shown, so the amplitude of firing is about the same for all delays. However, the response amplitude of this parietal neuron does depend on the direction and magnitude of the saccade, indicating that it is potentially capable of recording quantitative information about the task in its firing rate (Gnadt and Andersen, 1988). Neurons that show similar sustained firing patterns, associated with memory for specific modalities of information, are found in several cortical areas, for example, in IT cortex for vision (Fuster and Jervey, 1981, 1982; Fuster, 1990), posterior parietal cortex for touch (Koch and Fuster, 1989; Zhou and Fuster, 1992), and auditory cortex (Gottlieb et al., 1989).

Examples of failure to respond to stimulation when a memory task is not performed are shown in Figure 1, *B* and *C*. In Figure 1*B*, the animal performs an audition match-to-sample task only when a reward tube that delivers juice is in its mouth. The illustrated neuron shows a sustained response when a reward is anticipated and none when it is not, even though the same set of stimuli are presented. A frontal memory unit identified by its task sensitivity is shown in Figure 1*C*. Although the firing pattern of this neuron is contingent on the memory task, it is not a sustained activity unit. Rather, it fires briskly only during both the initial stimulus and final cue periods. This shows that other types of neurons, in addition to those with altered firing during the delay period, may be involved in the mechanism of information storage. A possible role for such neurons in the storage mechanism is suggested by the model described here.

The fact that similar types of memory-relevant neurons are found in different cortical areas suggests that the same kind of generic neural circuit, or module, is used to store active information throughout the cortex. If this is the case, then while the origin and significance of the stored information may differ from one cortical region to another, the kind of circuity used to store it may be the same. In this article we describe a neural network model of a circuit that can serve as such an active memory module.

Two broad classes of mechanisms have been frequently proposed to explain how neural firing is maintained during active short-term memory. In one, sustained firing is maintained by rapidly and temporarily changing synaptic strengths or other physiological parameters (Gottlieb et al., 1989). In the other, sustained firing depends only on neural activity recirculating in a network with fixed recurrent, that is, "re-entrant," connections (Cowan, 1972; Dehaene and Changeux, 1989; Zipser, 1991). The model described here was designed to test the hypothesis that networks with fixed, recurrent connections are sufficient to account for the observed experimental data.

In previous work (Zipser, 1991) it was shown that a simple model based on sustained activity in a neural network with fixed recurrent connections could account for part of the observed experimental data. This original model has some characteristic dynamical features, called fixed-point attractors, that should also be present in the dynamics of cortical activity if similar

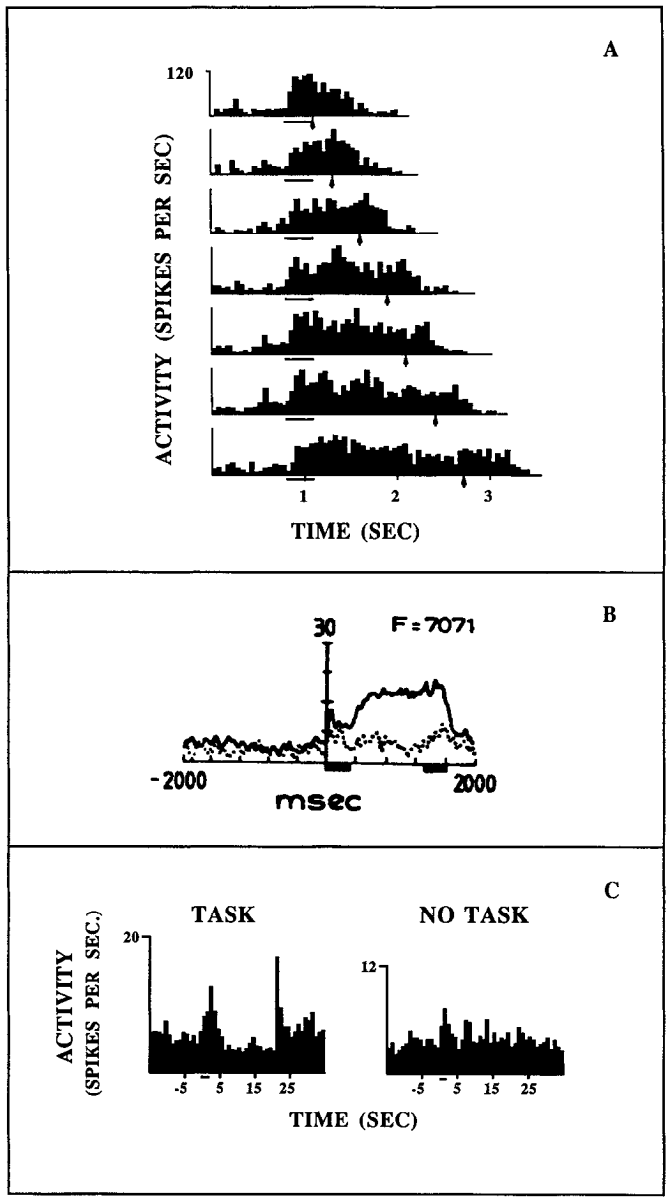



*Figure 1.* *A*, Spike histograms of an intended movement cell in area LIP of the rhesus monkey. Each histogram includes responses from 8–10 trials. Trials are grouped and ordered according to increasing response delay times. The *horizontal line below each histogram* indicates the stimulus presentation. The *arrow* indicates the time at which the fixation spot was extinguished. Eye movements occurred from 150 to 400 msec following offset of spot. Bin size = 50 msec. From Gnadt and Andersen (1988). *B*, Histogram showing the activity of a unit in the supratemporal gyrus of baboon auditory cortex during a tone matching task. *Dark bars* show the times of presentation of the first and second tones. *Solid line* is the task performance case, and *dotted line*, the no task case. From Gottlieb et al. (1989). *C*, Spike discharge histograms of a prefrontal unit during short-term memory performance. Bin size = 1 sec. The *short horizontal bar* indicates stimulus presentation. Red, green, yellow, or blue presented during the stimulus period indicate memory task with reward. Violet presented during the stimulus period indicates no prospective reward (no task cue). The neuron in *C* responds primarily to the initial stimulus and the final cue, about 20 sec later. From Yajeya et al. (1988) (figure from Zipser, 1991).

recirculating networks serve active memory there. Such attractor dynamics had originally been predicted to play a role in short-term memory by Cowan (1972). The presence of attractors would be expected to have significant effects on the spiking

patterns of cortical neurons in active memory circuits. However, because the original network used nonspiking model neurons, it could not be used to generate the spiking patterns needed to test for the effect of attractors. To overcome this difficulty, we have developed a more realistic, spiking version of the original model that enables us to compare predicted and observed spiking patterns. This provides a more valid test of whether the model is consistent with the experimental observations.

The spiking model described in this article was derived from the original model using a simple transformation that converts a neural network made up of continuous-output neurons into one with spiking neurons. The spiking model can load and store information in the form of neural spiking activity without the need to change any synaptic strengths during the process. It provides detailed information about the dynamic activity patterns and spiking statistics to be expected from recurrent active storage networks. This information can be compared directly to data obtained from single-unit recordings made in monkeys engaged in short-term memory tasks.

Here we first describe the original nonspiking neural network model and how it is transformed into a spiking model. We then examine the dynamic spiking patterns of model neurons and compare them to real single-unit data. Finally, we compare the statistical structure of long-term spiking patterns of real and model neurons to see if the real neurons have features indicative of fixed-point attractors. We find that the dynamic spiking patterns of many of the units in the model network closely resemble those of real neurons. Some real neurons also show spiking patterns that are predicted by the mechanism used in the model to load new information into memory. We also show that many real neurons involved in active memory do have spiking patterns indicative of the presence of fixed-point network attractors.

The model described here deals only with the circuit module used for the actual storage of active short-term information. Many other important issues about short-term memory, such as the neuroanatomy involved, the origin and mode of generation of the load signals, and the mechanism used to access stored information, are not addressed.

## Materials and Methods

The present study uses a set of 179 single-unit discharge records obtained in a previous study of the inferotemporal (IT) cortex of monkeys performing a visual delayed matching task (Fuster, 1990). Basically, the task required the animal to retain features (color or shape) of compound stimuli, each stimulus consisting of a colored disk (2.5 cm diameter) with a gray geometric symbol in the middle. On each trial, and depending on the symbol in the stimulus, the animal had to memorize—for 10–20 sec—either the symbol itself or the background color. Consequently, for correct performance of the task, each trial required attention to the symbol, in some trials also to the color, and short-term (10–20 sec) memory of either symbol or color.

The fully trained monkeys were surgically prepared for chronic microelectrode recording following procedures authorized by the UCLA School of Medicine (Division of Animal Medicine) and according to animal use guidelines from the National Institutes of Health and the Society for Neuroscience. All surgical operations were conducted with the animal under general anesthesia with Nembutal (slow intravenous infusion, about 35 mg/kg). The surgery essentially involved the implant of microelectrode carrier pedestals and head fixation bolts in the skull. Antibiotics were systematically and topically administered for prevention of infection. Head restraint during performance was gradually introduced to avoid discomfort. Test sessions lasted ordinarily some 3 hr, during which the animal consumed about 200 ml of liquid reinforcement.

The roving microelectrode used for extracellular unit recording during testing was made of platinum-iridium or Elgiloy and insulated with glass. Unit spike records were amplified and stored on computer disks. Only single-unit records, consisting of spike trains from isolated cells—judging from the voltage and shape of the spikes—are used in the data set studied here (see Fuster, 1990 for further details).

For purposes of spike analysis, the task was divided into the following periods: (1) baseline period of 15 sec preceding each trial; (2) stimulus presentation period; and (3) delay (10–20 sec feature memorization period). Statistically significant differences were determined using $t$ tests with 1% confidence limits unless otherwise noted in the text.

All the models described in this article were simulated on general purpose digital computer workstations using programs especially written for this project. Details about the models are described in the following section.

## Models of Active Memory

Artificial neural networks are being widely used to investigate the processing or computation carried out by networks of real neurons in various parts of the nervous system. The model neural units used in these networks are designed to approximate only the input–output properties of real neurons without attempting to simulate their inner workings. The outputs of these model neurons are generally represented by continuous values that can be compared to average spiking rates. These networks are simulated on digital computers to run in discrete time steps, rather than in continuous time. The process or computation carried out by a neural network is ultimately determined by the values of the synaptic weights that interconnect the neural units. These weights are generally chosen automatically by some optimization, or *training*, procedure. This training procedure is just a mathematical technique to pick a best set of weights, and is not likely to resemble the way this process occurs biologically. In spite of their simplicity, models of this type can often simulate the average spiking behavior of neurons in biological systems quite realistically (Zipser and Andersen, 1988). In the case of recurrent networks, the dynamics of the model often simulate the experimentally observed dynamics. This makes it possible to generate models with a close functional homology to biological systems. Particularly realistic results have been obtained recently with recurrent network models of the dynamics of the vestibulo-ocular system (Arnold and Robinson, 1989; Anastasio, 1991). The general paradigm for making artificial neural network models of the nervous system is called *neural system identification* (Zipser, 1992).

To model active short-term memory we tried to find the simplest network architecture that could account for the main features of the process. The available experimental data, particularly the dependence of memory-related activity on reward expectation, suggest that loading information into active memory is not an automatic consequence of the presence of a stimulus, but requires an additional, task-dependent loading signal. This implies that the simplest active memory module must have at least two inputs, one carrying the information to be stored and the other carrying a signal to indicate when that information should be loaded into storage. The information to be stored would, in general, be processed from an external stimulus and be specific for the modality of each cortical area. The load signal, on the other hand, is likely to be far less stimulus specific, and to be common to many memory modules in different cortical areas.

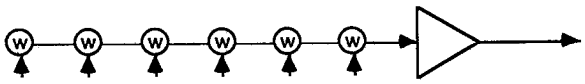
A previous model by Zipser (1991) has the required architecture and was used as the basis for the spiking model described here. We first describe this model and how it was trained to implement the short-term memory task. Then we show how

trained versions of the original nonspiking model are converted to spiking models.

*The original nonspiking model*

The original model consists of a recurrent network with an *information input,* a *load input,* and one output. The information input carries a continuous value representing the information to be stored. The load input carries a binary signal that is kept at zero as long as information is to be held in memory and set to 1.0 only when new information is loaded. The output carries the value of the stored information buffered from changes on the information input line.

The network consists of a set of recurrently connected model neurons, each of which can be represented schematically as follows:



where the large triangle represents the soma, the line to the left represents the dendrite and synapses, with inputs distributed along its length, and the arrow to the right represents the axon or output. The inputs to this neuron from other neurons in the network on time cycle, $t$, are represented by a set of activities, $y_j(t)$, with values equal to the outputs of all the $N$ units in the network. The model makes no assumption about the anatomical location of the various units in the network, but it is reasonable to assume that at least some of them are located near one another in the same cortical area. The external input sources to the network are represented by $z_k(t)$. In the active memory model there are two external inputs. One carries the information to be stored, and the other, the load signal indicating when to store new information. The strengths, or weights, of the connections between units in the network are represented by $w_{ij}$, where $i$ is the index of the postsynaptic unit and $j$ is the index of the presynaptic unit. The weights of the connections for external inputs are represented by $v_{ik}$. Each unit also has a bias, $b_i$, which is roughly equivalent to the sum of the resting potential and any unchanging afferent activity. The output of the $i$th model unit in the network on time cycle $t + 1$ is given by

$$y_i(t + 1) = f\left(\sum_j w_{ij}y_j(t) + \sum_k v_{ik}z_k(t) + b_i\right),$$

where $f(x)$ is the logistic function

$$f(x) = \frac{1}{1 + e^{-x}},$$

which ranges between 0 and 1 as $x$ ranges from minus infinity to plus infinity. The logistic function is a sigmoid, which roughly captures the observation that real neuron firing rates cannot be less than zero and have some maximum value.

The full network is presented schematically in the top panel of Figure 2. The schematic diagram shows all units interconnected and all receiving external inputs. It illustrates the potential for interconnection before training. After training some weights may go to zero, so in the actual memory model not all possible connections are functionally present. Note that only one unit in the network carries the output value. The other units are called *hidden units.* They carry out the processes involved in the memory task; they are the model units whose properties are compared to real neurons.

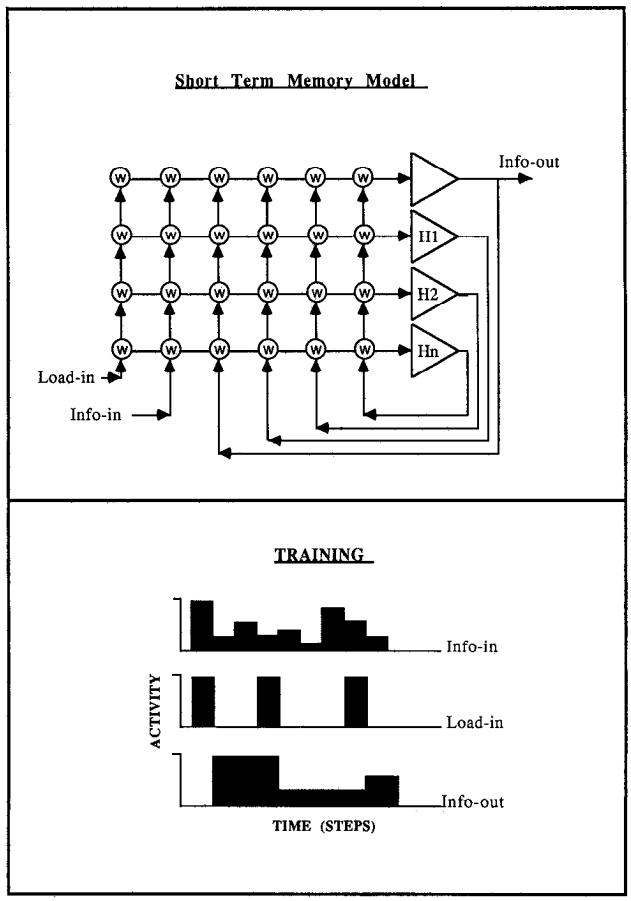



*Figure 2.* The structure of the model. *Top,* Input and recurrent connections. *Bottom,* Diagram of the training paradigm.

The memory task the network was trained to do is illustrated in the lower panel of Figure 2. A gradient-descent, error correction optimization algorithm for recurrent networks, called *Backpropagation Through Time* (Williams and Zipser, in press), was used to find weight values that would allow the network to implement the task. Networks were initialized with random weight values between −1.0 and 1.0. These values were adjusted by the optimization algorithm on each cycle of training until the network output matched the target required by the task to within an error of ±5%. The bias quantities, $b_i$, were not adjusted but fixed at −2.5 to guarantee that unstimulated units have low activities. On each step of training the network was given an *Info-in* input value chosen randomly between 0.0 and 1.0 and a *Load-in* input value of 0.0 (off), except during randomly chosen cycles when the *Load-in* signal was set to 1.0 (on) to load a new value. The average time between load signals was four time steps. The output of the network (*Memory-out*) was trained to maintain the value of *Info-in* at the time the *Load-in* signal was on. Typically, 50,000 training cycles were required to reach the error criterion. Many instances of this model were trained with sizes ranging from 6 to 20 units.

Zipser (1991) showed that hidden units in networks trained in this way exhibited many of the properties of cortical neurons relevant to short-term memory. We used these trained nonspiking networks as the basis for our spiking model in the manner described below.

*Converting continuous models to spiking models*

The basic idea used to convert nonspiking models to spiking models was to interpret the continuous output of the nonspiking model neurons as spiking probabilities (Cowan, 1968; Amit, 1990). This approach gets considerable justification from both early and recent studies of the statistics of cortical neuron firing (Fuster et al., 1965; Smith and Smith, 1965; Sejnowski, 1976; Softky and Koch, 1992; Snowden et al., 1992). Further support is provided by a recent study of the spiking characteristics of some of the IT neurons used in this study (Littlewort et al., 1992).

Our goal in making spiking networks was to have them behave as much as possible like the trained continuous networks from which they derived, while at the same time incorporating enough realism to allow comparison with experimental data. Each individual spiking neuron computes its probability of spiking in the same general way a continuous neuron computes its output value. Spiking neurons, however, produced output values of 1 with this probability, and output values of 0 with 1 minus this probability. In this discrete-time model of spiking, each time step in which a spike occurs can be considered to be the combination of a single spike and the absolute dead time that follows it.

The dynamic activity of a recurrent network cannot be maintained by simply replacing each continuous unit with a spiking one. The binary valued outputs together with a low firing rate would completely disrupt the orderly function of the network. This problem was overcome by replacing each continuous unit in the original network with a pool of many spiking neurons. The average number of spikes produced by a pool is proportional to the output activity of the neuron it replaced. The problem of maintaining recurrent activity with slowly spiking neurons is not limited to model networks but also arises in the nervous system (Amit, 1990). The spiking network consists of the same number of pools of spiking units as the original network had nonspiking units. Each member of a pool was connected to neurons in other pools the same way as the neuron it replaced, but with the weights appropriately scaled to take into account the number of units in each pool. There are no connections between units in a pool. The spiking behavior of the individual neurons in these pools is assumed to be comparable to that of single neurons in the brain.

The equation governing the behavior of the spiking unit is

$$y_i(t+1) = 1 \quad \text{with probability } sf\left(\sum_j w'_{ij}y_j(t) + \sum_k v_{ik}z_k(t) + b_i\right),$$

$$y_i(t+1) = 0 \quad \text{with probability } 1 - sf\left(\sum_j w'_{ij}y_j(t) + \sum_k v_{ik}z_k(t) + b_i\right),$$

$$0 < s < 1.$$

The scale factor $s$ serves to keep the spiking rate of the model neurons low and comparable to that of real neurons. The recurrent weight values used in the spiking model, $w'_{ij}$, are those of the continuous model divided by $ns$ to compensate for both the $n$-fold increase in number of inputs and the scaling of the output by $s$. The range of the index $i$ is now from 1 to $nN$, where $N$ is the number of neurons in the continuous model and $n$ is the size of each spiking pool. Note that the output of a neuron, $y_i$ now takes on only the values 0 or 1 with a probability given by the logistic function scaled by $s$.

Neurons involved in short-term memory have a wide range of average firing rates, but none show persistent firing anywhere near their physically maximum rate. This maximum rate is sometimes observed when a neuron is accidentally injured. We simply use the scale factor, $s$, to allow simulated neurons operating near saturation to have the possibility of arbitrarily low spiking rates. This is of importance to the model because without $s < 1$ some model neurons would have firing probabilities near 1, leading to unrealistic ceiling effects on spiking statistics. In our formulation, all neurons in the model use the same value of $s$. For values of $s < 1$ the spiking version of a model is no longer exactly homologous to the continuous version. The reason for this difference is subtle and has to do with changes in spiking variance introduced by removing the ceiling effects. In practice these effects are small enough that the spiking version still behaves essentially like the original model.

## Behavior of Short-Term Memory Model Networks

In this section we describe and explain the behavior of the model itself. The next section compares the model with experiment. First, the firing patterns of model neurons for periods that approximate those of short-term memory experiments are described. Then the long-time behavior of the network is analyzed to show the relaxation to fixed-point attractors and how it is affected by random spiking. Finally, we explain the general mechanism used by the model to load and store information.

*Short-term dynamic behavior*

The typical short-term memory experiment consists of an intertrial baseline period, a brief stimulus presentation, a delay during which information about the stimulus is remembered, and finally, another stimulus presentation to which the animal responds based on remembered information. Our model, however, is concerned only with the loading and storage of information, and not with the specific modality of the information or how it is accessed and used to respond. We assume that our model represents a short-term memory module somewhere in the cortex. During simulated memory experiments it receives stimulus-relevant afferent information on its *Info-in* line and load signals during stimulus presentations on its *Load-in* line. During the intertrial interval and the delay period a fixed baseline value is held on the *Info-in* line. During stimulus periods a value to be stored, representing the effects of the stimulus on this particular module, is put on the *Info-in* line while the *Load-in* line is set to 1.0. The load line is reset to 0.0 during the delay. The dynamic activation patterns of model neurons produced by this paradigm for one instance of both the original nonspiking and the spiking model are shown in Figure 3. (Fig. 3 shows a model instance called f11, which is used as an illustrative example throughout this article). A relatively short memory delay period, commensurate with those used in training, is used here. Each unit has its own characteristic activity pattern. The output unit has a moderately stable sustained activity reflecting the stored value. This shows that the network has learned to approximate the memory task on which it was trained. The spectrum of observed hidden unit activity patterns can be roughly divided into three major classes: units with sustained activity during the delay differing from baseline, that is, hidden units 1, 2, and 5; units with major activity changes only during the stimulus period, that is, hidden units 3 and 6; and more complex

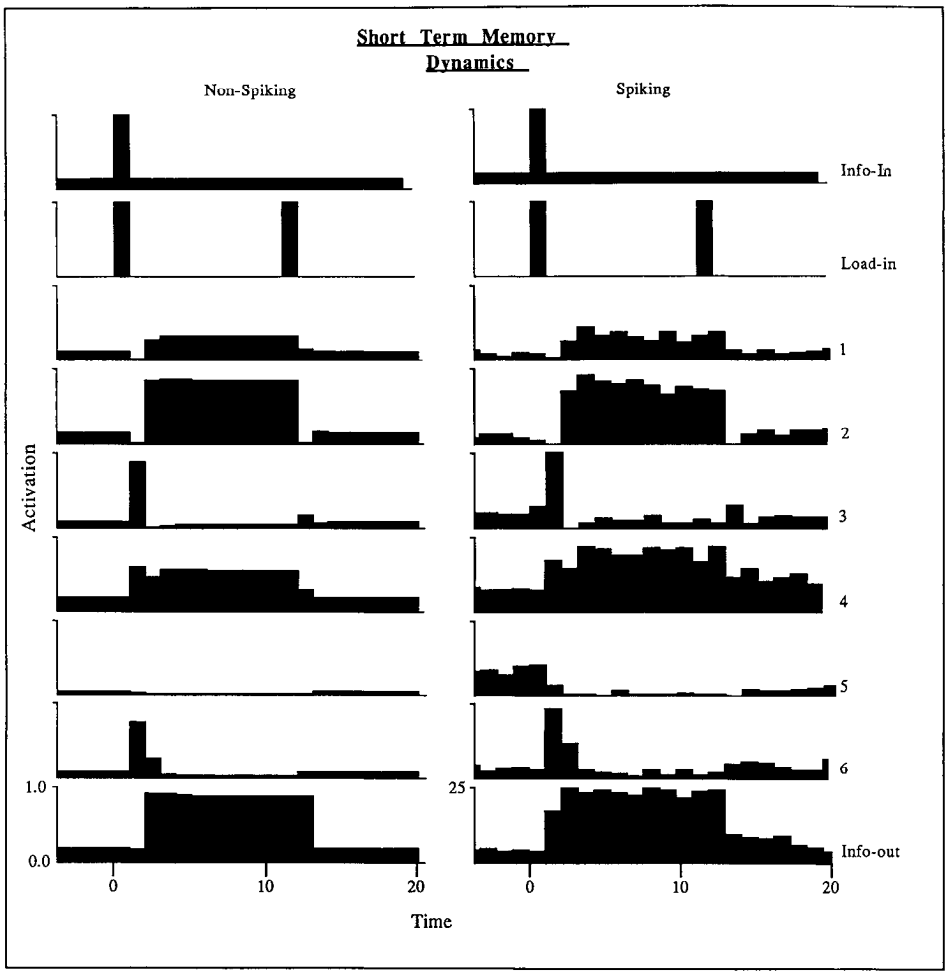



*Figure 3.* Temporal activity patterns of units in the continuous and spiking version of model instance fl1. The continuous network consisted of six hidden units and one output unit. It was trained to implement a system with the input–output characteristics described in Figure 2 and the text. The bias weights are fixed at −2.5 in this model instance. Training was for 200,000 time steps with an average of four time steps between load pulses. The patterns for the continuous version, on the *left,* were generated by first setting the activities to their basal levels by loading in an *Info-in* value of 0.1 (the load signal is not shown). Then, a value of 1.0 is loaded in and held for nine time steps, during which time the input is held at 0.1. Then, with the input still at 0.1 the load signal is given again. The patterns for the spiking version on the *right* were generated in the same way. The spiking version has 160 neurons per pool and an *s* value of 0.3. The histograms were made by collecting the spikes from one neuron in each pool on 500 trials. The *vertical axis* is spikes per 100 trials per time step.

units that mix the previous two characteristics, that is, hidden unit 4. Note that unit 5 shows sustained inhibition. Other instances of the model generated in different training runs and starting with different sets of initial random weights all have hidden units with these three basic classes of activity patterns, but differing in detail. The role played by the different classes of hidden units in the mechanism of storage will be discussed later.

*Long-term dynamics and attractors*

When the delay period is greatly lengthened, the original, nonspiking network can no longer maintain its stored value, but relaxes to a stable state called a fixed-point attractor. This is a consequence of the fact that recurrent networks with the nonlinear units used here cannot store arbitrary values indefinitely. Cowan (1972) demonstrated attractor dynamics for recurrent networks of nonlinear units, and suggested that the fixed-point attractors might play a role in short-term memory. Other kinds of attractor states such as limit cycles or chaos are theoretically possible, but all of the original networks displayed only fixed-point attractors. In the majority of cases there were two stable states for the network, with rare examples of one or three (Zipser, 1991).

In the stable state the outputs of all units in the nonspikiing network remain constant as long as there is no change on the external inputs. The stable state to which a network will finally settle is determined by a distinct input value that serves as a kind of threshold; for inputs below this threshold the network moves in time to one attractor, and for inputs above it the network settles to the other. This is illustrated in Figure 4. Note that while sustained-activity hidden units 1, 2, and 4 go to high states for the above-threshold attractor, unit 5 goes to a low state. This situation is reversed for the below-threshold attractor. The threshold value and low state for unit 5 are so close they are not resolved in Figure 4. Also note that the high and low attractor activity values reached by the sustained-activity units are not the extremes of their possible range, that is, 0.0 and 1.0, and are quite different for each unit. Each instance of the model has its own characteristic threshold and settling time. These attractor states behave somewhat differently in the spiking model, and are critical for understanding the experimentally observed spiking statistics.

*Attractor dynamics of the spiking model*

The introduction of random spiking into the model creates a source of noise, and noise disrupts the stability of the fixed-point attractor states (Cowan, 1972; Zipser, 1991). After an initial period during which the stored value is approximately maintained, noisy networks do not stay permanently in a single attractor state but move, at random times, from one attractor to another, spending most of the time near an attractor, but a significant amount of time in transition as well.

The movement between attractor states in the long-term dynamics of a spiking model is shown in Figure 5. The memory

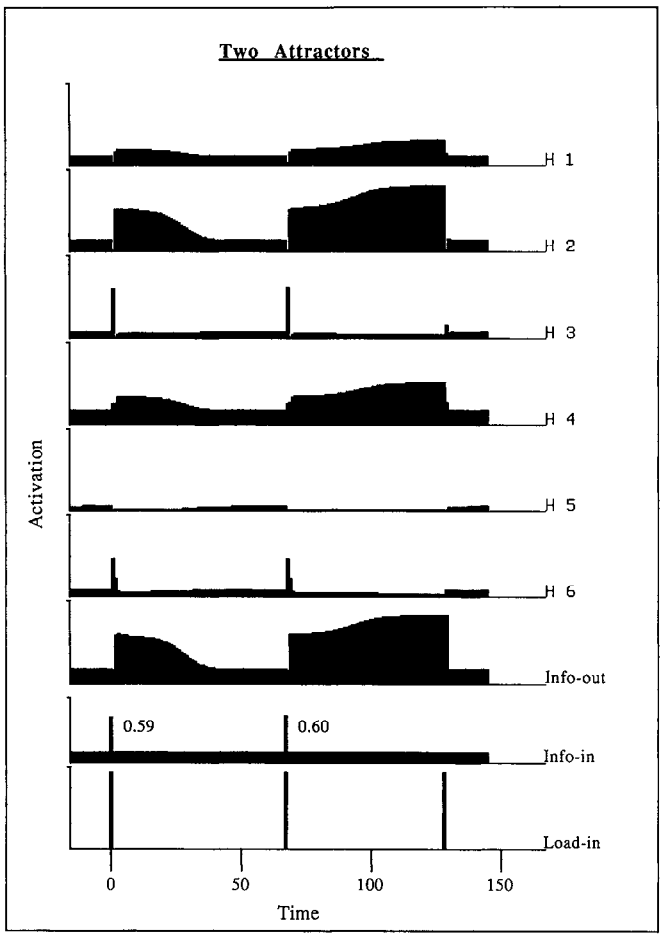



*Figure 4.* Attractor dynamics of the continuous version of model instance f11. The threshold for the model instance shown here is 0.595. When values above this threshold are loaded into the network, all units settle to their upper attractors; when values below the threshold are stored, they settle to their lower attractors. The figure shows graphs of the temporal activity patterns obtained for a pair of starting values just below and just above the threshold. The time course of activity in the network is displayed for 60 time steps between load signals.

was loaded with a high value, after which the input was held constant for more than 1800 time steps while the activity of each unit was sampled every four steps. The total number of spikes produced by each pool on the sampled time steps is used as a measure of activity. Note that the activity of each pool starts off near its appropriate level for a high stored value, and then switches randomly, spending most of the time near one of its attractor values. Some time is also spent at intermediate values. The length of time spent near any attractor is quite variable, ranging from a few to nearly 100 steps. Note also that the behavior of all the sustained-activity units, for example, 1, 2, 4, and 5, is highly correlated, indicating that the attractor states are features of the network as a whole and not assignable to individual units.

### *How the model works*

The general strategy used by the model to load and retain information can be determined by examination of the connection weights and the activity patterns of the units. By the end of training the output unit has become functionally a separate layer since all its weights from the input lines and all its feedback weights to the rest of the network have become nearly zero. Most of the other neural units in a model become either *storage* or *gating* units. Storage units sustain an activity representative of the stored value during the delay. The gating units have high activity during the stimulus periods and near-baseline activities otherwise. Storage units act as a group to maintain their sustained activity through shared recurrent feedback connections. The storage units are of two basic kinds, *positive* and *negative.* Outputs of the positive storage units monotonically increase as the stored value increases. Outputs of negative storage units decrease as the stored value increases. Both the positive and negative units feed back onto all other units of the same kind with excitatory weights, and onto units of the opposite kind with inhibitory weights. Once a value is established in the storage units, it tends to persist, at least for the effective storage time of the memory. However, each storage unit represents the stored value with its own characteristic activity level. The storage units receive inhibitory input from the *Load-in* and either excitatory or inhibitory input from the *Info-in* line. This seems to help in resetting the memory before a new value is stored.

The gating units receive excitatory input from both the *Load-in* and the *Info-in* and project excitatory output to all the positive storage units and inhibitory output to all the negative storage units. This allows the gating units to pass a new memory value into the storage units on the time step after the *Load-in* signal goes back to zero. The gate units receive weak mixed recurrent connections that have little net affect. This means that during the delay they are dominated by the strong negative bias that all units have keeping their activities near their baseline. This tends to buffer the storage units from activity changes on the *Load-in* line.

Some units are more complex than those described and seem to combine features of both storage and gating units. As we shall see, all major classes of units found in the model correspond to neurons found in the cortex. It is also interesting to note that, if negative storage units are left out, it should be possible to have model networks in which all the recurrent connections are excitatory. This actually seems to be the case, since one of us (B.K.) has successfully trained instances of the original model constrained to have only excitatory recurrent connections. One of the two input signals still has to be inhibitory, but nonetheless such a network obeys Dale's law, that is, that the outputs of any given neuron are all of the same polarity.

## Model versus Experiment

The validity of a model can be tested only by comparing its behavior to that of the real system in as many ways as possible. For practical reasons we are currently limited to comparing the behavior of single units in the model to single-unit firing patterns recorded from the brain, primarily from IT cortex in the case of this article. Two different aspects of these firing patterns are compared. First, the average firing patterns produced during real and simulated short-term memory tasks are compared to show that many real and model neurons have similar response properties. Then, certain statistical properties of firing during the intertrial baseline period are compared to show that real neurons behave as if they are in noisy networks with fixed-point attractors.

### *The average firing patterns of real and model neurons*

In this section we show that individual neurons in the spiking model behave like real neurons during short-term active memory experiments. First, we compare the peristimulus histograms

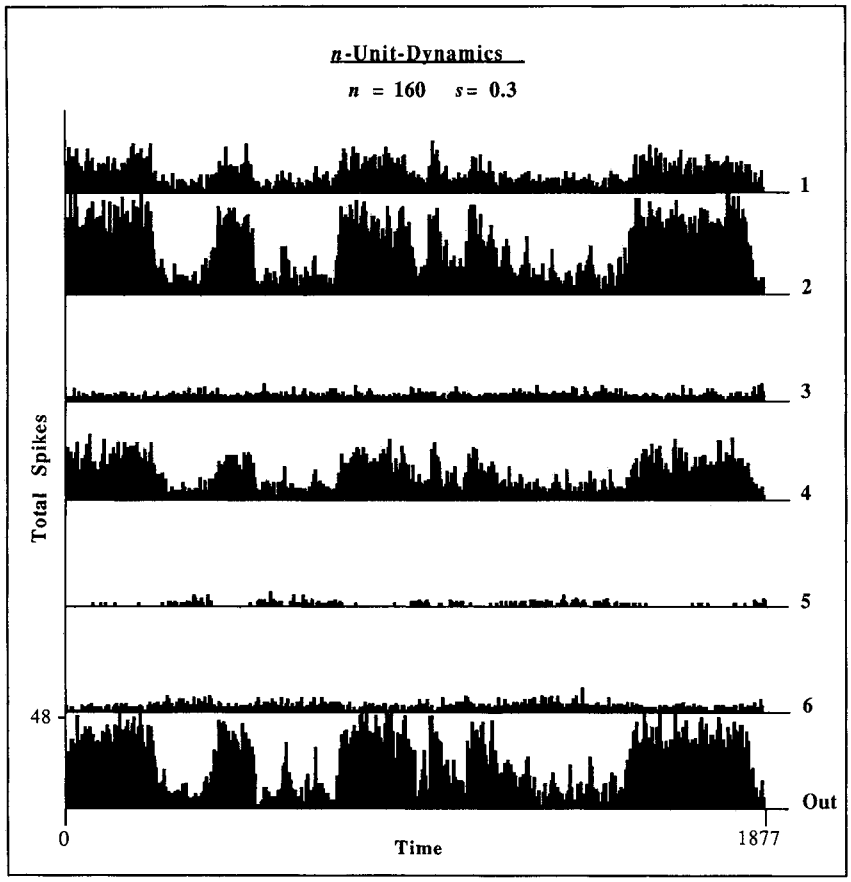



*Figure 5.* Long-term temporal activity patterns of the spiking version of instance f11. The model was run as in Figure 4, but after gating in a value of 1.0 no further gating was done. The activity for about 1800 cycles is shown as total spikes per pool. The activity is sampled only every four cycles.

recorded during short-term memory experiments typical of cortical neurons to simulated histograms from selected spiking model neurons. This demonstrates that many of the patterns of activity typical of cortical memory neurons are also found in model neurons. Then, we look more closely at the way real and model neurons respond to different stimulus properties. This comparison shows the detailed homology between the differential response of real and model neurons. Finally, we deal with the question of what fraction of real and model neurons are actually comparable. This analysis shows that virtually all the model neurons have homologs among cortical neurons, and that a considerable fraction of the memory-relevant cortical neurons are accounted for by the model.

Biological short-term memory experiments were simulated on trained networks by loading an information value, representing the stimulus to be remembered, and holding it for the delay period by keeping the load signal at zero. During the delay period and during the intertrial intervals an *Info-in* value, different from the stimulus, was used to represent the background input level present in the absence of a stimulus. To make the model and real peristimulus histograms comparable, certain parameters not determined by the model or known from the experimental data must be fixed. For example, we don't know the constant that relates model time and real time, and we also have no way of measuring the actual level of the afferent signals being stored in real memory. Because these parameters must be chosen arbitrarily, only comparisons between the shapes of the activity patterns of real and model neurons are relevant.

Examples of typical real and model firing patterns during active memory tasks are shown in Figure 6. Close examination of Figure 6 shows that several characteristic features of the model's activity patterns are found in the experimental data. For example, the sustained activity units in the model differ as to how they respond when the load signal is active. This difference is mirrored in the experimental units shown in Figure 6, *A* and *B*. Note that the real neuron in Figure 6*B* is inhibited during the periods when the initial and final stimuli are being presented. This corresponds directly to the model unit where the inhibition is caused by the reset process that occurs when new information is loaded into memory. Another feature found in both model units and real neurons is the tendency of the sustained activity to drift up or down during the delay period, as seen in Figure 6, *C* and *D*. This has previously been attributed by one of us either to a decay of the stored information or to anticipation of the upcoming action (Fuster, 1984, 1989). In the case of the model these changes are due to the the network moving toward fixed-point attractors.

A limitation of the kind of comparison shown in Figure 6 is that it is based on the response to only a single value of afferent stimulation. A more compelling comparison can be made if the differential responses of real and model neurons are compared. Many neurons in the IT data set used here showed differential responses to different stimuli, that is, red versus green, + versus o. Presumably this is the result of different values being stored to represent the different stimuli. The differential response of two kinds of model neurons, a storage unit and a gate unit, were compared with two corresponding IT neurons (see Fig. 7*A,B*). The peristimulus histograms of the real neurons are generated from all trials on which either red or green was the initial stimulus. The differential responses for the model neurons are gen-

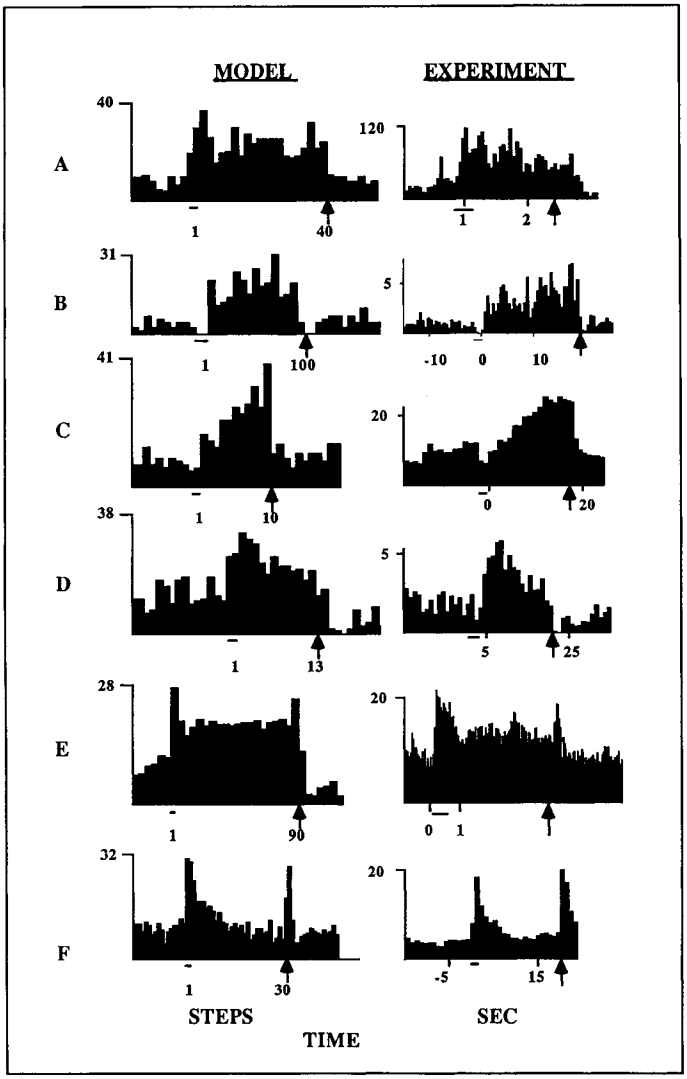



*Figure 6.* Comparison of the temporal activity patterns of cortical neurons with hidden units from spiking model networks during real and simulated delay memory experiments. The experimental data have been copied from published sources using a Hewlett Packard ScanJet Plus. The histograms have been redrawn to the same physical size and format to facilitate comparison. The model data come from different units in five independently trained instances of the original continuous model to which the spiking network transformation was applied. The *horizontal axis* represents time, in seconds for the real neurons and in time steps for the model units. The *vertical axis* represents activity in spikes per second for the real neurons and spikes per 100 trials per time step for the model units. The spiking models all had 500 units per pool and an $s$ value of 0.3. The *horizontal bar* is the time of presentation of the first stimulus in the case of the experimental data and indicates the period immediately after the offset of the first load in the case of the model. The *arrow* indicates the start of the cue ending the delay period in the experimental case and the offset of the final load in the case of the model. *A* is a neuron from posterior parietal area LIP during a delay saccade task, from Gnadt and Andersen (1988). *B* is an IT neuron during a visual delay match-to-sample task, from Fuster et al. (1985). *C* and *F* are frontal neurons in the principal sulcus during delay match-to-sample experiments, from Fuster (1984). *D* is a frontal neuron during a delay choice task experiment, from Quintana et al. (1989). *E* is a composite of 33 principal sulcus neurons that all have cue-period and delay-period activity in a delay saccade task, from Funahashi et al. (1990).

erated by storing a high value to simulate the response to one color and a low to simulate the response to the other. Many trials are run, accumulating spikes from a single model neuron in a pool. These data are then used to generate *high* and *low* peristimulus histograms for comparison with the red versus green response of real neurons.

The differential activity of IT23.27B, in Figure 7*A*, closely corresponds to that of storage-type hidden unit 2 of the model instance f11 shown in Figure 3. They are both strongly inhibited during all stimulus periods. During the delay, the red response of IT23.27B mimics the response of the model unit when a high value is stored in memory, and its green response mimics the model unit's low storage value response. The differential activity of IT23.40A, in Figure 7*B*, closely corresponds to that of gate unit 3, also from the model instance f11 shown in Figure 3. Both IT23.40A and the model unit have above-baseline activation only during the stimulus period.

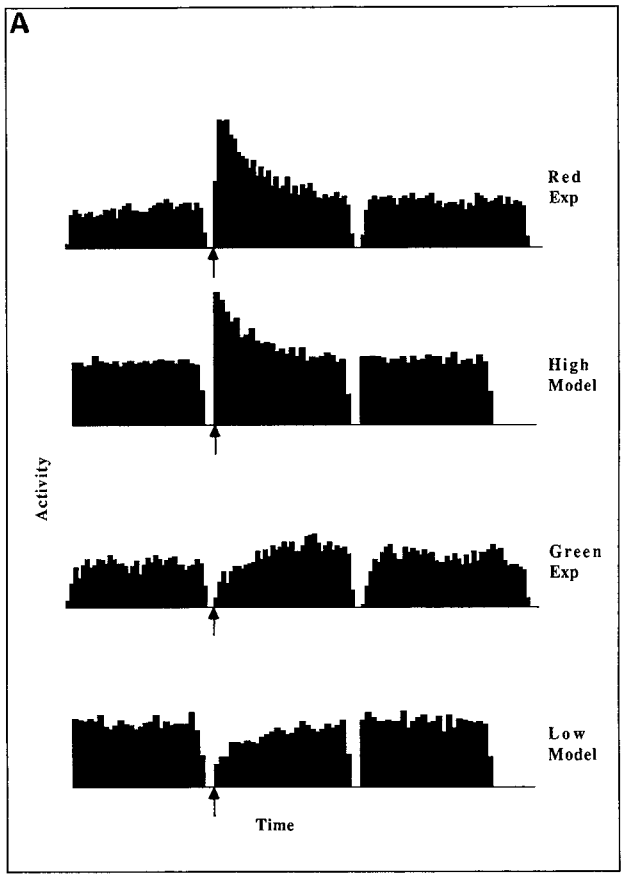
The kind of similarities between model and real neurons seen in Figure 7 are not rare observations. Similarities occur with some regularity in both the model and the brain. To quantify this evidence we compared the neuron types in a set of 179 memory-relevant, single-unit recordings from IT with 48 hidden units from eight instances of the spiking model, all independently trained with the same parameters as instance f11, but with different random values of the starting weights. We tried to find objective tests to decide if two neurons were of the same type. Global comparisons of this kind are difficult because there are many different types of model and real neurons, and the behavior of both model and real neurons depends on the afferent value being stored. We used two approaches to this problem. One approach was to match model and real neurons on the basis of qualitative features, independent of any differential responses to stimuli. The other was to categorize real and model neurons on the basis of their differential responses. The first approach was applied only to units of the two types illustrated in Figure 7. The second approach was used on all neurons with a significant differential response to red and green.

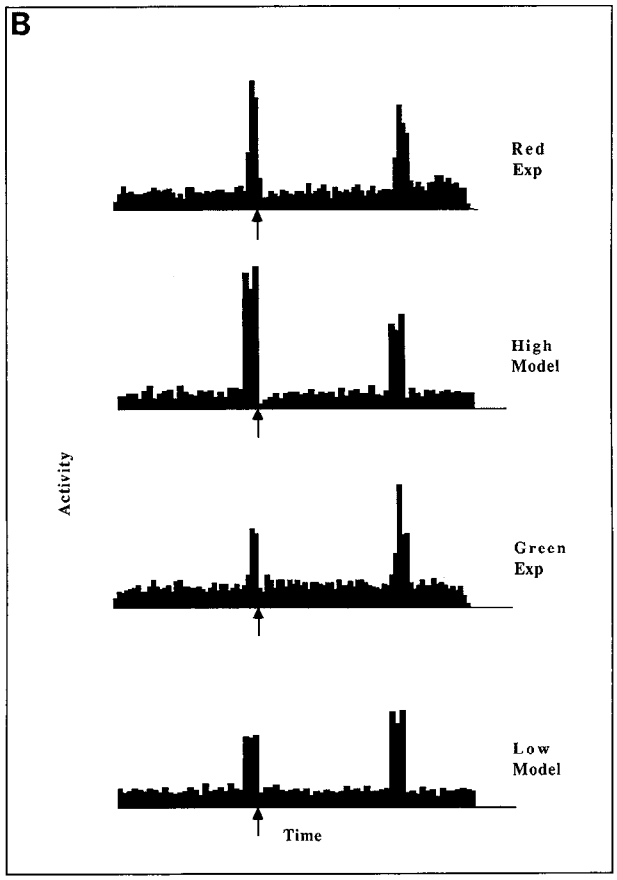
To find out how many real and model units matched the types illustrated in Figure 7, we chose criteria that would enable the computer to search for units that matched these two prototypes. The criteria used to identify model units of the same type as unit 2 of model instance f11 were the following: significant inhibition below baseline during the stimulus period, and a systematic increase in activity with stored value during the delay. The criteria used to find real units of the same type as IT23.27B were the same as for the model during the stimulus period. However, since we have no way of systematically varying the afferent input to the real neurons, the constraint imposed on activity during the delay was that it changed at the start of the delay and moved monotonically toward the baseline, or remained constant during the delay. With these criteria, 17 (9.5%) of the real neurons and 15 (31%) of the model neurons were of the type of IT23.27B or unit 2, respectively. Only two of the real neurons had highly significant differential responses during the delay period. The rest had the same response during the delay to both red and green. Of these nondifferential units, nine were similar on both kinds of stimulation to either the red response or the green response of IT23.27B. The rest simply returned quickly to the baseline level as soon as the stimulus was removed and remained there for the rest of the delay. All of these observed delay responses found in the real neurons were consistent with them being of the same type as IT23.27B, as-

suming an appropriate afferent value was present during stimulation. Some examples of units classified as of the same type as IT23.27B are shown in Figure 8*A*.

The criteria used to identify model units of the same type as gate unit 3 of instance fl1 were excitation in the stimulus period that increased systematically with the *Info-in* value, and near-baseline activation throughout the delay period. For neurons to be considered of the same type as IT23.40A, they had to have significantly above-baseline activation in the stimulus period, together with near-baseline activation throughout the delay period. With these selection criteria, 48 (27%) of the real and 9 (19%) of the model units were of the same type as unit 3 of fl1. Three examples are shown in Figure 8*B*. A significant differential response to red and green was found in 10 of the real neurons of the same type as IT23.40A. Together, the two different types of unit illustrated in Figure 8 represent 40% of the model units and 36.5% of the real neurons.

Another approach we used for global matching of real and model units was based on their differential responses to afferent stimulation. This analysis was confined to the 50 real neurons that showed a significant differential response to red or green during either the first stimulus or the delay, and to the 44 model units that showed significant differential responses to high or low stimulation. Each unit was characterized as being above or below baseline during the stimulus, and above, below, or within 10% of baseline during the delay period. Thus, for example, IT23.27B was classified as a "− +/− −" type because it was below baseline for both stimulus periods and above baseline for red and below it for green during the delays. IT23.40B was of type "+ 0/+ 0" since it had greater-than-baseline responses in both the stimulus periods which were significantly different from each other, and was very near baseline during the delay periods. Note that this is only a qualitative categorization and is applied only to units that have significant differential responses. The order of the two parts of the categorization is irrelevant because we don't know the afferent values coming from red or green stimulation; we only know their effects. If we knew the relative magnitudes of these values there might be many more types. Only trials on which red or green were paired with the "="




→

*Figure 7.* *A* and *B*, Differential response of real neurons and spiking model neurons. These histograms cover five time periods: (1) 15 sec of prestimulus time in which the monkey is presumably alerted to the impending trial but has no information about the stimulus, (2) about 1 sec period during which one of two stimuli is presented, (3) an 18 sec delay with no stimulus present, (4) the response period during which both stimuli are present, and (5) about 20 sec of postresponse activity. The peristimulus histograms of the model neurons have the same five periods. *A*, *Red Exp* and *Green Exp* show the response of neuron IT23.27B to red and green initial stimuli, respectively; *High Model* and *Low Model* show the response of unit 2 of model instance fl1 with 50 units in each pool and $s = 0.3$ to gating in a high, 0.99, or a low, 0.01, initial stimulus, respectively. The model data were accumulated from a single unit on 1000 trials. The *first arrow* indicates the time the stimulus went off in experimental cases and the time the load signal was turned off in model cases. The *second arrow* indicates the time the response cue came on in experimental cases and the time the load signal started for a second time in model cases. There was an 18 sec period between the two arrows in the experiment and 53 cycles of model time in the model. The bin size for the experiment is 500 msec, and for the model it is two cycles. The load signal on time for the model was five cycles. *B* is the same, except that in it IT23.40A is compared to unit 3 of model instance fl1, and the low value stored is 0.3.

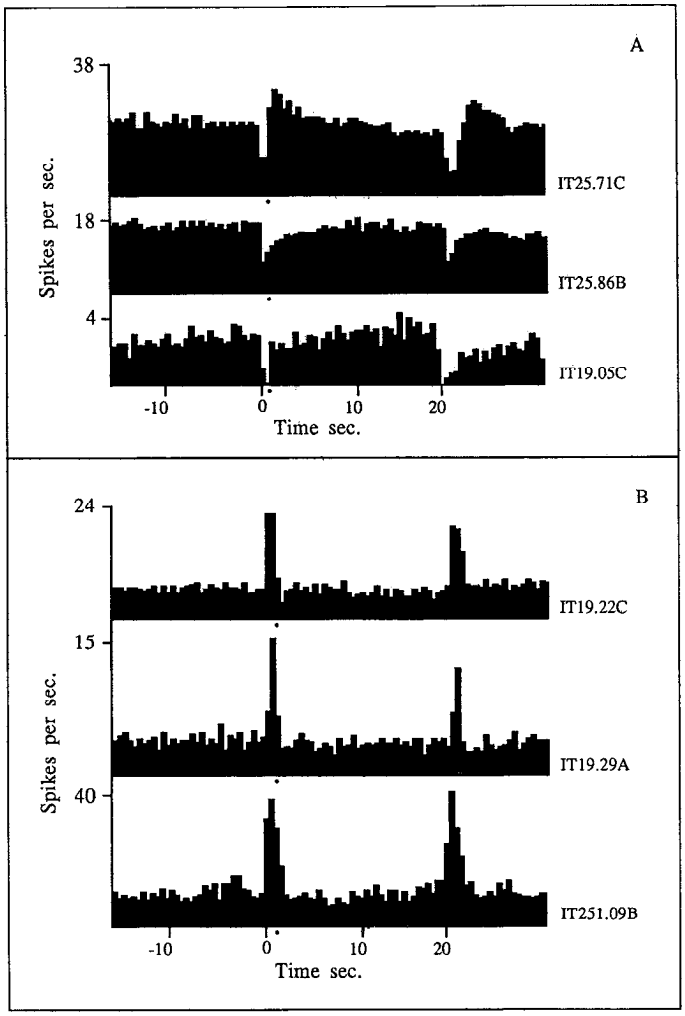



*Figure 8.* *A* and *B*, Examples of units of the same type as IT23.27B and IT23.40A. *A* shows three examples of IT neurons that do not have red–green differential responses, but are otherwise of the same type as IT23.27B. *B* is the same, but here the neurons are of the same type as IT23.40A.

symbol, indicating color to be relevant (Fuster, 1990), were used, because with our simple model we have no unambiguous way to simulate a condition in which two, possibly interacting, differential stimuli are present. Only the first 5 sec of the delay were used for comparison because the tendency to decay back to the baseline for most of the delay period can mask significant differential responses. Of the 11 types considered, all 50 real neurons fell into 10 and the 44 differential model neurons fell into 7. All 7 of the model types were a subset of the 10 real neuron types. The data are shown in Table 1. Out of the 50 real differential response neurons, 41 were of types found in the model.

Taken together, the data from both global matching procedures show that all the differential response patterns of model neurons can be matched, at least qualitatively, with real IT neurons, and that a significant fraction of the real neurons can be matched with model neurons. Since we have not attempted to match all nondifferential-response real neurons, the degree of matching may be underestimated. We would not expect all IT neurons to match model neurons since it is unlikely that the model accounts for all kinds of processing going on in IT cortex.

**Table 1. Spiking model and IT differential-response neurons of each type**

| Type | Number in model | Number in IT |
|---|---|---|
| + 0/+ 0 | 9 | 10 |
| + −/+ − | 0 | 5 |
| + +/+ + | 1 | 9 |
| − −/− − | 0 | 2 |
| − +/− + | 0 | 2 |
| + −/+ + | 10 | 7 |
| + +/− + | 2 | 4 |
| − +/− − | 15 | 2 |
| + −/− − | 1 | 7 |
| + +/− − | 6 | 2 |
| Total | 44 | 50 |

All model and IT neurons were first tested for a significant difference between low and high or red–green response, respectively, in either the stimulus or the delay period. If there was a significant difference they were then classified into types on the basis of whether their responses were above, "+," or below, "−," baseline, except in one case where a class consisted of units with a delay response within 10% of baseline, "0." The "/" separates one stimulus–delay pair from the other, but the order is irrelevant since we don't know for the IT neurons whether green or red gives the larger afferent signal.

## Looking for Attractors

Fixed-point attractors are a characteristic feature of our model. If networks like those of the model are present in the cortex, then cortical neurons should also exhibit attractor dynamics. One way to get evidence for attractor dynamics in the cortex is to detect the characteristic temporal structure attractors impose on spike trains. This structure consists of time segments with discrete spiking probabilities corresponding to each of the attractors, together with segments with changing probabilities generated while the neuron is moving between attractors. Detecting these segments in data from single neurons is difficult because the segments are of varying lengths, and interspike intervals of all sizes can occur in any given segment. Only the average spiking rate differs between segments.

One way to visualize the structure imposed by network attractor dynamics is to look at the running average of the spiking rate. Figure 9 provides examples of actual spike trains and their running averages from the spiking model and real neurons that show the kind of temporal structure imposed by noisy attractor dynamics. The model data are taken from a network that has been running long enough that it is no longer affected by its starting state. According to our model, its firing should represent what is seen in single neurons contained in networks randomly switching between fixed-point attractors. The experimental data on IT neurons are taken from the 15 sec intertrial baseline period and thus represent the "background" level of firing. Examination of Figure 9 shows that there are segments of fast and slow firing that extend over significant periods of time in both model and real spike trains. In the model this structure is due to attractors. Since no memory task is being performed during the baseline period, it is reasonable to assume that the memory-relevant IT neurons are not being driven by afferent, stimulus-related signals or receiving-load signals. Thus, the major changes observed in their firing rates could be the result of internally generated processes such as those associated with attractor dynamics. Direct examination of spike trains in this way shows the existence of structure in the real data that is similar to that found in the model. However, this technique is not compelling

because it requires subjective comparisons between very noisy signals. Stronger conclusions require techniques that are more objective and take all the available data into account in making comparisons.

*Multiple interval histograms*

One way to detect structure in spike trains is to use multiple-interval histograms (Gerstein and Mandlebrot, 1964; Tuckwell, 1988). Adding together several interspike intervals tends to average out the random lengths of the individual intervals and give a better measure of local spiking probabilities. The distribution of these multiple interval lengths will tend to have a mode for each discrete persistent spiking frequency. These models will be visible only in favorable cases where the temporal persistence of discrete states is sufficiently long. Even when the individual modes are not visible, the existence of temporal structure can be detected in multiple-interval histograms by comparing them to histograms of the same multiplicity made from randomly shuffled data. Randomly shuffling the order of spike intervals destroys all temporal structure. The effect of this on the shape of the shuffled distribution depends on the kind of structure present originally. For example, if the original spike train consists entirely of a long interval separated by a very short interval, then the shuffled distribution will be broader than the original because shuffling generates runs of short and long intervals that did not exist originally. This will lead to a shuffled distribution with a greater variance than the original distribution. However, if the original train is rich in runs of fixed spiking probabilities, shuffling will narrow the distribution by destroying long runs of pure high and low spiking rates. This will produce a reduction in the variance of the shuffled compared to the unshuffled distribution. These considerations suggest that multiple-interval histograms could be used to detect attractor-like structure in spike trains. We have used this technique to compare model and experimental data.

The properties of a multiple interval histogram depend on the number of consecutive intervals, $m$, used. If $m$ is too small it cannot average out the effects of random interspike times and the distribution will not capture interesting temporal structure. If $m$ is too large it will cover so much of the spike train that all interesting structure is averaged out. In practice our results were not very sensitive to the value of $m$. Significant differences between the shuffled and unshuffled distributions, when present, persist over a wide range of $m$ values. Many IT neurons show large significant differences between the original and the shuffled distributions that persist over ranges of $m$ from eight to 32 intervals. These effects are of the magnitude and kind expected from the effects of fixed-point attractors in the spiking model. An example is shown in Figure 10, where multiple-interval histograms are compared for a real IT neuron and unit 2 of model instance f11. In both cases the shuffled distribution is narrower, with fewer high- and low-frequency intervals and a lower variance. The unshuffled distributions show indications of more than one mode that disappear in the shuffled distributions.

Are these variance shifts significant, and if so, how many neurons show them? If the shuffled and unshuffled distributions really have different variances, then the probability of finding a shuffled distribution with the same variance as the original unshuffled one should be very small. To test this we shuffled each spike train 100 times and found the mean and SD of the 100 resulting variances. We then measured the difference between the variance of the unshuffled distribution and the mean of the 100 shuffled distributions in units of the SD of the variance of the shuffled distributions. The SD is a convenient measure of this difference because it is time-scale invariant and gives an indication of both the significance and magnitude of the difference.

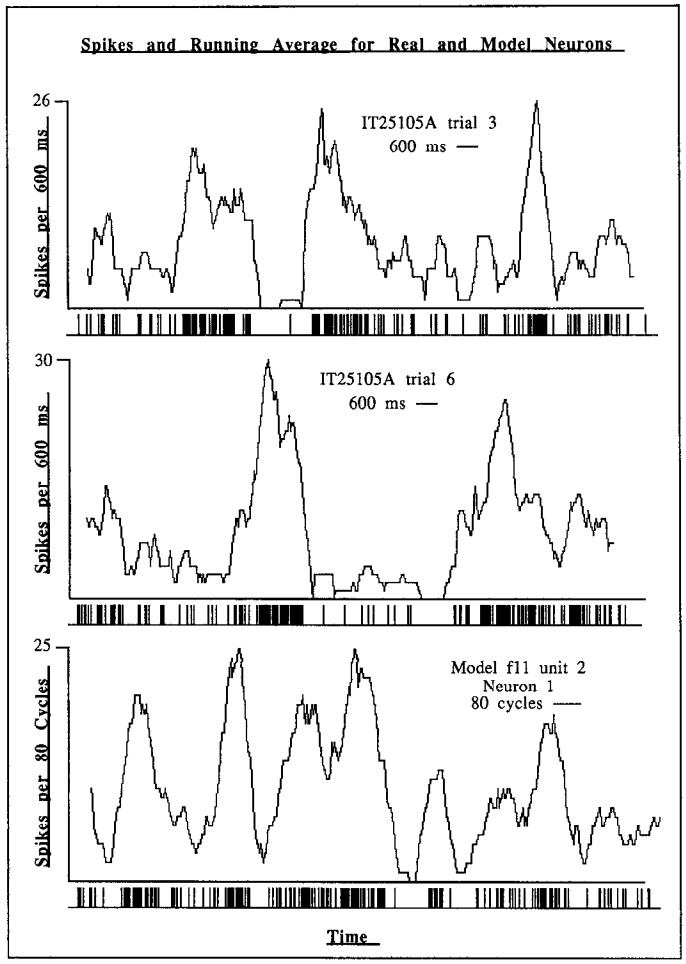



*Figure 9.* Spike train and running average from a real and a model neuron. The running average is generated by summing all the spikes in a fixed-width time window and plotting this number for each millisecond or cycle. The bin size is 600 msec for the real neurons, 80 cycles for the model. The scale of the spike train is too coarse to resolve individual spikes closer than about 30 msec.

We first looked for variance shifts in a selected sample of 20 single-unit spike trains that had high firing rates, very short refractory periods, and "exponential"-like single-spike interval distributions. For the set of 20 selected neurons the average difference between the variances of the shuffled and unshuffled distributions was 25 SD for $m = 8$, 31 SD for $m = 16$, and 34 SD for $m = 32$. The corresponding values for unit 2 of model instance f11 are 38 SD for $m = 8$, 37 SD for $m = 16$, and 24 SD for $m = 32$. These differences are very significant since the probability of finding so many SDs difference at random is vanishingly small. The effect is quite insensitive to $m$, at least over the range of 8–32 intervals. All the neurons in our sample of 20 had significant differences between shuffled and unshuffled distributions; the range was 7–77 SDs.

We then measured the variance differences for all the neurons in our IT data set. Of the 179 neurons tested, the differences found between the variance of shuffled and unshuffled multiple-

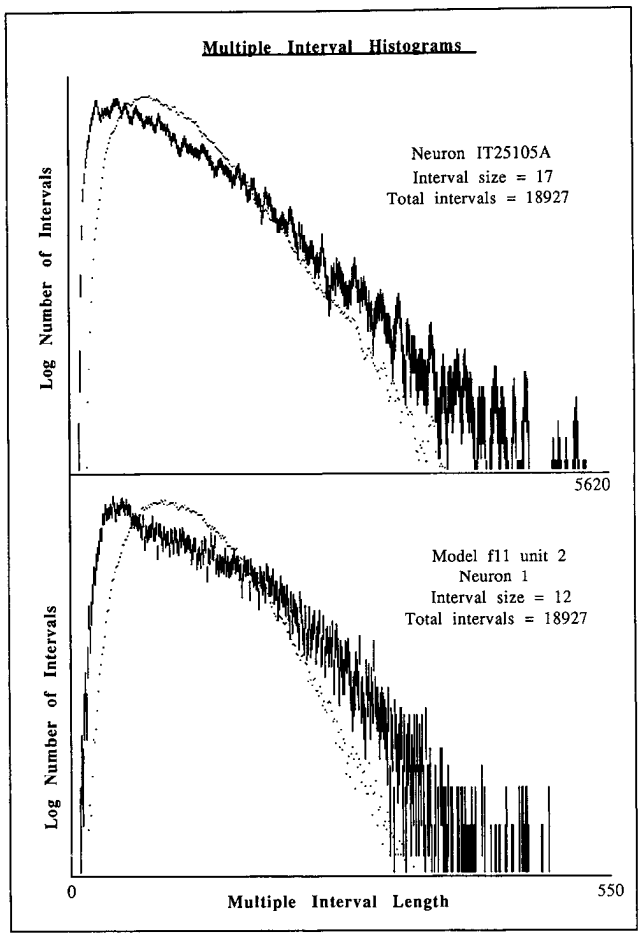



*Figure 10.* Multiple-interval histograms. Multiple-interval distributions are generated by summing the time for $n$ spikes to occur after each spike, and then binning these times. Number of intervals summed = 15 for IT251.05A and 12 for f11. The same procedure is applied to the original data and to shuffled data created by randomly reordering all interspike intervals. Because any given random shuffling can have some idiosyncratic features, the shuffled distributions are the average of 10 randomizations. The unshuffled distributions are plotted using error bars centered on the number of intervals with lengths of two times the square root of the number of intervals. The shuffled distributions are plotted as points without error bars. The model and neuron plots have been scaled to have about the same overall size. The IT251.05A graph has been smoothed by averaging three adjacent intervals at each time point.

interval distributions were less than 55 SDs for 43%, between 5 and 10 SDs for 16%, and greater than 10 SDs for 41%. These data represent the largest difference found for $m$ values of 8, 16, and 32. Some examples of the the range of values obtained are given in Table 2. Note the uniformity over $m$ values. Only three neurons out of 179 have differences between distributions that span the range of $<5$ SDs to $>10$ SDs as $m$ goes from 8 to 32. If we assume that neurons with more than 10 SDs difference in the variance of shuffled and unshuffled distributions have extensive temporal structure in their spike trains, then these observations are consistent with the assumption that more than 40% of the neurons in IT are in networks with noisy fixed-point attractors.

## Discussion and Conclusions

The spiking model investigated here uses very simple probabilistic neurons. These model neurons are "realistic" in the sense

**Table 2. Examples of differences in variance between shuffled and unshuffled multiple-interval distributions**

| Neuron | Multiple interval size | $V_u - V_s$/SD |
|---|---|---|
| IT19.03A | 8 | 0.45 |
| | 16 | 0.42 |
| | 32 | −0.16 |
| IT19.15A | 8 | 12.6 |
| | 16 | 11.8 |
| | 32 | 10.2 |
| IT23.13A | 8 | 35.0 |
| | 16 | 43.2 |
| | 32 | 55.5 |
| IT25.83B | 8 | 51.8 |
| | 16 | 66.2 |
| | 32 | 77.0 |
| IT19.21A | 8 | −0.06 |
| | 16 | 7.6 |
| | 32 | 16.0 |
| IT251.05A | 8 | 38.4 |
| | 16 | 37.3 |
| | 32 | 24.0 |
| Unit 2 f11 | 8 | 38.2 |
| | 16 | 42.6 |
| | 32 | 44.4 |

$V_u$ is the variance of the unshuffled multiple-interval distribution. $V_s$ is the mean of the variances of 100 shuffled distributions. SD is standard deviation of the 100 shuffled variances.

that their output spiking statistics are similar to those observed in many IT neurons. We cannot say if their input–output characteristics are realistic because little is known about the *in situ* transfer statistics of neurons in the cortex (Softky and Koch, 1992). No attempt is made to derive the behavior of the model neurons using basic knowledge about cellular properties; that is, it is a black box model of individual neurons. Model networks can show realistic behavior to the extent that their model neurons capture the important input–output statistics of neurons in the cortex. These networks are useful in analyzing brain function because the major features of phenomena such as loading information into active memory and the structure imposed on spiking by attractors are likely to be fairly independent of the fine details of neuron function. In the cases analyzed here both the average temporal dynamics and the statistical properties of the model spike trains correspond to what was found in many real neurons thought to be involved in short-term active memory. This demonstrates that the model is consistent with the experimental observations. While this qualifies the model as a possible way to account for the observed data, it does not rule out alternative mechanisms for active storage. For example, it might be possible to also construct a model that uses rapid weight changes to account for the experimental observations. To our knowledge no one has done this yet. An alternative explanation for the temporal structure found in the spike trains, which our model attributes to noisy attractors, is that the units in IT are being driven by visual input as the monkey looks around the dimly lit room when not being directly stimulated. One line of evidence against this hypothesis is the finding that IT discharge is not correlated with eye movements during either the baseline or memorization periods (Fuster and Jervey, 1982).

Perhaps the two aspects of the model with the most significant consequences are the global signal that is used to load new information into active memory and the statistical flipping between spiking probabilities caused by the presence of fixed-point attractors. The model can shed relatively little light on the anatomical origin of the load signal or the chemical nature of the transmitter involved. Both must be common to many memory circuits because it is highly unlikely that any single, small, unreliable network could account for reliable memory behavior. It would be interesting if the transmitter mediating the load signal was associated with one of the neuromodulatory systems that innervate the cortex. Whether or not the hypothesized load signal exists, together with its anatomical location and mechanism of generation, can ultimately be determined only by additional experimental studies.

The results described here show that the spiking of many IT neurons is consistent with the notion that they belong to networks with fixed-point attractors disturbed by spiking noise. This is of interest not only for the present model, but also because attractors in spiking networks have been hypothesized to play a role in the representation of entities in long-term memory and in recall into active memory (Cowan, 1972; Amit, 1990). Long-term memory for discrete entities is hypothesized to be represented by the values of synaptic weights that determined the set of possible attractors. These networks have individual neurons with only two discrete firing probabilities. The networks have many attractor states, each with a different combination of neurons in the upper and lower state. Active short-term recall is implemented when the system settles into one of these attractors. This differs from our model, which is a storage mechanism for graded values. These values are generally represented by the network not being in an attractor state and are lost after settling into attractors. However, all suggestions concerning the use of attractors require that they exist, so our findings may be relevant to other models as well. Currently there is not enough experimental information to tell whether the attractor states we have evidence for in IT are associated with graded storage as in our model, discrete recall, or both kinds of memory.

Further confirmation of the validity of the model can come from multiunit recordings, which could demonstrate the predicted correlations between the activity of neurons in the same recurrent memory network. In the course of the original experimental work (Fuster, 1990) some multiunit recordings were made, but were not included in the data set used in this study. We have examined some of these multiunit recordings and find that in several cases they have the same statistical patterns expected of single units in attractor networks. While not enough detailed information is available from these multiunit recordings to determine for sure that all units are in the same network, these observations suggest that it may eventually be possible to trace out the hypothesized recurrent networks in more detail using multiple-unit recording.

The model described here consists of a neural network module designed to account for a specific body of well-established experimental data about sustained firing observed in very simple short-term memory tasks. The model deals only with the issue of immediate active information storage and is not a complete model of short-term and working memory. Many important issues that have been addressed experimentally, such as memories maintained across multiple rapid stimulus presentations, the mechanism of matching current and remembered stimuli (Miller et al., 1991), and active recall of learned pattern stimuli (Sakai and Miyashita, 1991), are not addressed by our model. However, we can find nothing in the experimental findings on these issues that rules out the use of active storage modules of the type described here for information storage over short periods of time. Indeed, the striking similarity between the spiking patterns predicted by the model and those found in IT neurons demonstrates that a mechanism based solely on recurrent connections that do not change rapidly can account for the data on *sustained* activity observed during short-term memory. This result in no way conflicts with the likelihood that there is also rapid synaptic weight change, for example, in the hippocampus, needed for the full range of short-term memory phenomena.

It is important to note that this consistency between model and experiment was not directly designed into the model, but emerged indirectly from the process of optimizing a network of nonlinear neural-like units to implement a simple memory task. This result is another example of the power of applying the systems identification paradigm to neural networks for generating models of nervous system function (Zipser, 1992).

## References


Amit DJ (1990) Attractor neural networks and biological reality: associative memory and learning. Future Generation Comput Sys 6:111–119.

Anastasio T (1991) Neural network models of velocity storage in the horizontal vestibulo-ocular reflex. Biol Cybern 64:187–196.

Arnold D, Robinson DA (1989) A learning neural-network model of the oculomotor integrator. Soc Neurosci Abstr 15:1049.

Bauer RH, Fuster JM (1976) Delayed-matching and delayed-response deficit from cooling dorsolateral prefrontal cortex in monkeys. J Comp Physiol Psychol 3:293–302.

Cowan JD (1968) Statistical mechanics of nervous nets. In: Neural networks: proceedings of the school on neural networks, June 1967 in Ravello (Caianiello R, ed), pp 181–188. Berlin: Springer.

Cowan JD (1972) Stochastic models of neuroelectric activity. In: Statistical mechanics (Ricce SA, Fread KF, Light JC, eds), pp 181–182. Chicago: University of Chicago.

Dehaene S, Changeux J (1989) A simple model of prefrontal cortex function in delayed-response tasks. J Cognit Neurosci 1:3.

Funahashi S, Bruce CJ, Goldman-Rakic PS (1990) Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. J Neurophysiol 63:814–831.

Fuster JM (1973) Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. J Neurophysiol 36:61–78.

Fuster JM (1984) Behavioral electrophysiology of the prefrontal cortex. Trends Neurosci 7:408–414.

Fuster JM (1985) The prefrontal cortex, mediator of cross-temporal contingencies. Hum Neurobiol 4:169–179.

Fuster JM (1989) The prefrontal cortex: anatomy, physiology, and neuropsychology of the frontal lobe, 2d ed. New York: Raven.

Fuster JM (1990) Inferotemporal units in selective visual attention and short-term memory. J Neurophysiol 64:681–697.

Fuster JM, Alexander GE (1971) Neuron activity related to short-term memory. Science 173:652–654.

Fuster JM, Jervey JP (1981) Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science 212: 952–955.

Fuster JM, Jervey JP (1982) Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. J Neurosci 2:361–375.

Fuster JM, Herz A, Creutzfeldt OD (1965) Interval analysis of cell discharge in spontaneous and optically modulated activity in the visual system. Arch Ital Biol 103:159–177.

Fuster JM, Bauer RH, Jervey JP (1982) Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks. Exp Neurol 77:679–694.

Fuster JM, Bauer RH, Jervey JP (1985) Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. Brain Res 330:299–307.

Gerstein G, Mandlebrot B (1964) Random walk models for the spike activity of a single neuron. Biophys J 4:41–68.

Gnadt JW, Andersen RA (1988) Memory related motor planning activity in posterior parietal cortex of macaque. Exp Brain Res 70: 216–220.

Goldman-Rakic PS (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In: Handbook of physiology: the nervous system (Mountcastle VB, Plum F, eds), pp 373–417. Bethesda, MD: American Physiological Society.

Gottlieb Y, Vaadia E, Abeles M (1989) Single unit activity in the auditory cortex of a monkey performing a short-term memory task. Exp Brain Res 74:139–148.

Koch KW, Fuster JM (1989) Unit activity in monkey parietal cortex related to haptic perception and temporary memory. Exp Brain Res 76:292–306.

Littlewort GC, Zipser D, Perez P, Fuster JM (1992) A spiking attractor network model accounts for bimodal statistics of visual cortex neurons. Soc Neurosci Abstr 18:740.

Miller EK, Li L, Desimone R (1991) A neural mechanism for working and recognition memory in inferior temporal cortex. Science 254: 1377–1379.

Niki H (1974) Prefrontal unit activity during delayed alternation in the monkey. Brain Res 68:185–204.

Quintana J, Yajeya J, Fuster JM (1988) Prefrontal representation of stimulus attributes during delay tasks. I. Unit activity in cross-temporal integration of sensory and sensory-motor information. Brain Res 474:211–221.

Quintana J, Fuster JM, Yajeya J (1989) Effects of cooling parietal cortex on prefrontal units in delay tasks. Brain Res 503:100–110.

Sakai K, Miyashita Y (1991) Neural organization for the long-term memory of paired associates. Nature 354:152–155.

Sejnowski TJ (1976) On the stochastic dynamics of neuronal interaction. Biol Cybern 22:203–211.

Smith DR, Smith GK (1965) A statistical analysis of the continual activity of single cortical neurones in the cat unanesthetized isolated forebrain. Biophys J 5:47–74.

Snowden RJ, Treue S, Andersen RA (1992) The response of neurons in areas V1 and MT of the alert rhesus monkey to moving random dot patterns. Exp Brain Res 88:389–400.

Softky WR, Koch C (1992) Cortical cells should fire regularly, but do not. Neural Comput 4:643–646.

Tuckwell HC (1988) Introduction to theoretical neurobiology, Vol 2. Cambridge: Cambridge UP.

Williams RJ, Zipser D (in press) Gradient-based learning algorithms for recurrent networks. In: Back-propagation: theory, architectures, and applications (Chauvin Y, Rumelhart DE, eds), in press. Hillsdale, NJ: Erlbaum.

Yajeya J, Quintana J, Fuster JM (1988) Prefrontal representation of stimulus attributes during a delay task. II. The role of behavioral significance. Brain Res 474:222–230.

Zhou Y, Fuster JM (1992) Unit discharge in monkey's parietal cortex during perception and mnemonic retention of tactile features. Soc Neurosci Abstr 18:706.

Zipser D (1991) Recurrent network model of the neural mechanism of short-term active memory. Neural Comput 3:179–193.

Zipser D (1992) Identification models of the nervous system. Neuroscience 47:853–862.

Zipser D, Andersen RA (1988) A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature 331:679–684.