Role of Inferior Temporal Neurons in Visual Memory

II. Multiplying Temporal Waveforms Related to Vision and Memory

EMAD N. ESKANDAR, LANCE M. OPTICAN, AND BARRY J. RICHMOND

Laboratory of Sensorimotor Research, National Eye Institute, and Laboratory of Neuropsychology.
National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892

SUMMARY AND CONCLUSIONS

1. In the companion paper we reported on the activity of neurons in the inferior temporal (IT) cortex during a sequential pattern matching task. In this task a sample stimulus was followed by a test stimulus that was either a match or a nonmatch. Many of the neurons encoded information about the patterns of both current and previous stimuli in the temporal modulation of their responses.

2. A simple information processing model of visual memory can be formed with just four steps: 1) encode the current stimulus; 2) recall the code of a remembered stimulus; 3) compare the two codes; 4) and decide whether they are similar or different. The analysis presented in the first paper suggested that some IT neurons were performing the comparison step of visual memory.

3. We propose that IT neurons participate in the comparison of temporal waveforms related to vision and memory by multiplying them together. This product could form the basis of a crosscorrelation-based comparison.

4. We tested our hypothesis by fitting a simple multiplicative model to data from IT neurons. The model generated waveforms in separate memory and visual channels. The waveforms arising from the two channels were then multiplied on a point by point basis to yield the output waveform. The model was fitted to the actual neuronal data by a gradient descent method to find the best fit waveforms that also had the lowest total energy.

5. The multiplicative model fit the neuronal responses quite well. The multiplicative model made consistently better predictions of the actual response waveforms than did an additive model. Furthermore, the fit was better when the actual relationship between the responses and the sample and test stimuli were preserved than when that relationship was randomized.

6. We infer from the superior fit of the multiplicative model that IT neurons are multiplying temporally modulated waveforms arising from separate visual and memory systems in the comparison step of visual memory.

INTRODUCTION

The neuronal mechanisms necessary to perform even the simplest cognitive tasks remain unknown. One basic task related to visual memory is sequential image comparison, or recency memory. As described in the companion paper, we studied monkeys trained to perform delayed nonmatch-to-sample tasks in which two pictures were briefly presented 550 ms apart. We found that the responses of inferior temporal (IT) neurons encoded information about the patterns of both current and previous pictures. This led us to hypothesize that some of the neurons in IT cortex are involved in comparing the internal representations of current and recalled stimuli. This paper tests a hypothetical neuronal mechanism for making this comparison.

The process of determining whether two pictures presented sequentially are the same or different can be decomposed into four steps (Fig. 1). In one step, the test image must be encoded by neurons in the visual system. In another step, an encoded representation of the sample image must be recalled from recent memory. Then, the two encoded representations, one from vision and one from memory, must be compared. Finally, the similarity of the two representations must be judged. How the brain performs each of these processes is not known.

It is commonly thought that pictures are represented in the brain by the distribution of activity across an ensemble of neurons. Furthermore, it is also believed that pattern recognition involves some form of distributed associative memory in which memory traces are stored in the weights of the connections between elements (Carpenter and Grossberg 1988; Grossberg 1976a,b; Kohonen et al. 1989; Rolls 1987). These hypotheses are based on static population codes and do not address the temporal modulation observed in neuronal responses. No mechanisms based on the temporally modulated neuronal activity observed during sequential pattern recognition have been proposed yet. Thus, many questions of how the brain actually performs recency memory tasks remain a mystery.

Although the mechanism of sequential comparison is not known, its site is likely to be in IT cortex. IT cortex is intimately related to earlier visual areas (Baizer et al. 1991; Desimone et al. 1980; Ungerleider and Mishkin 1987) and to limbic structures important for memory (Aggleton et al. 1980; Amaral and Price 1984; Insausti et al. 1987; Iwai and Yukie 1987; Iwai et al. 1987; Turner et al. 1980; Van Hoesen and Pandya 1975a,b; Webster et al. 1991; Whitlock and Nauta 1956). Lesion studies have established that the inferior temporal cortex is necessary for the performance of pattern recognition tasks that require vision and memory (Dean 1978; Iwai and Mishkin 1968, Mishkin 1972, 1982). Furthermore, single unit recordings have shown that the responses of IT neurons to visual stimuli are modulated in a variety of ways during visual memory tasks (Baylis and Rolls 1987; Eskandar et al. 1992; Fuster and Jervey 1981a,b; Gross et al. 1979; Mikami and Kubota 1980; Miller et al. 1991; Miyashita and Chang 1988; Moran and Desimone 1985; Riches et al. 1991; Richmond and Sato 1987; Spitzer and Richmond 1991; Spitzer et al. 1988). Thus IT cortex is a likely candidate for linking vision and memory.
The results in the companion paper showed that the responses of IT neurons encode information about both current and remembered stimuli. Those findings suggested that IT neurons might be involved in comparing visual stimuli with recalled stimuli. However, information theory is model-free and thus gives no indication of how such a comparison might be made. Therefore, a model of the comparison step must be based on inferences about neuronal mechanisms drawn from other results.

It has been shown that neurons throughout the visual system convey information with the use of temporally modulated waveforms (Cattaneo et al. 1981; Eckhorn and Pöppl 1974, 1975; McIlrkin et al. 1991; Optican and Richmond 1987; Richmond and Optican 1990). We have proposed that this temporal modulation plays a ubiquitous role in visual processing. In a visual memory task, this temporally encoded visual information must be compared with remembered information. This comparison can only be made if the visual and remembered information have the same internal representations. However, the way in which memories are represented in neural systems is not yet known. It is possible that visual information is converted into the same representation used by memory. Alternately, the remembered information may be converted into a temporally modulated representation matching that of the visual information. The advantage of converting memories into a temporal code is that a simple mechanism, such as crosscorrelation, could be used to make the comparison between current and remembered representations. These ideas lead us to propose a general hypothesis: that visually elicited temporal waveforms are compared with temporal waveforms recalled from memory.

The process of crosscorrelation can be divided into two steps: multiplication of two waveforms, and integration of the product. Identical waveforms will have a large integral and dissimilar waveforms a small one. If temporal modulation is indeed involved in the comparison of visual and remembered information, then it might be possible to detect the consequences of this process in the responses of IT neurons. For example, the activity of IT neurons may contain evidence for a multiplicative interaction between current and previous stimuli.

In this paper we test a specific hypothesis about the functional mechanism of the comparison step in visual memory: that neurons in IT cortex multiply waveforms arising from memory with those arising from vision. The product of such a multiplication could then be used in the comparison of the two waveforms. We examined these hypotheses by fitting a simple multiplicative model to data from IT neurons. An abstract of this work has been presented (Optican et al. 1991).

METHODS

Experimental data

We tested our model on data from 21 IT cortical neurons of one monkey during a sequential nonmatching-to-sample task with either three or six Walsh patterns, as described in the companion paper (Eskandar et al. 1992). Briefly, after the monkey fixated a visual target, the target went off and a sample picture was shown in the center of the monkey's visual field for 352 ms. After this picture went off, the fixation target reappeared for 550 ms and then went off again. A test picture was then shown for 352 ms. The test picture was randomly chosen as either a match or a nonmatch to the sample picture. We analyzed the responses of IT neurons over an interval starting 90 ms after stimulus appearance and lasting 256 ms. The neuronal spike trains were converted to continuous waveforms estimating the probability of spike occurrence. For our modeling, the continuous spike density waveforms were sampled at 4 ms intervals to yield discrete response waveforms defined at 64 points.

Hypothetical models

We tested the hypothesis that visually elicited temporal waveforms are multiplied with temporal waveforms recalled from memory by constructing a specific model of the proposed processing performed by IT neurons (APPENDIX). The model was structured as a three-layer network (Fig. 2) but not the conventional fully connected neural network used by connectionists (Rumelhart et al. 1986). We hypothesize that this network models the multiplicative interaction that occurs between temporal waveforms representing visual and remembered information.

The first two layers of our network were divided into independent channels, one for the previous and one for the current picture. The two channels of the input layer received the pictures encoded as Walsh transforms. The components of the input Walsh transforms were all zero except for one which was unity, because the pictures were themselves selected from the Walsh basis set. The two channels of the input layer were composed of either three or six units, corresponding to the number of Walsh patterns used in a given experiment.

The second, or hidden, layer also had two channels, each of which contained 64 units, corresponding to the 64 sample points of the temporally modulated waveforms. These units were nonlinear, having the sigmoidal function (ranging from 0 to 1) standard in many neural network models (APPENDIX, Eq. 46) (Rumelhart et al. 1986). The input neurons in each channel were fully connected through adaptive weights to each of the hidden units within the same channel. The purpose of this structure was to convert the input pictures into a temporal response waveform 64 samples long. We interpreted the internal waveforms generated by the model as waveforms that might arise from vision and from memory.

The third layer consisted of a single channel of 64 units with linear activation functions. Each output unit corresponded to one sample of a temporal response waveform. The output units had paired inputs, one from each of the two channels. Thus the two inputs to the first unit in the output layer were from the first unit in the memory channel and from the first unit in the visual channel, and so on. In this way the output waveforms were causally
The model was trained by adjusting the weights between the input and the hidden layers, for each of the two channels, and by adjusting the biases on the output units. The training set was formed by pairing the sample and test Walsh stimuli with the recorded neuronal responses to the test stimulus. The training set consisted of the responses from individual trials. The models were trained, with the use of the gradient descent method with weight decay (see APPENDIX), until the change in mean-squared error (mse) calculated from the models' predictions of all the waveforms reached a fixed convergence criterion. The weight decay ensures that the model uses internal waveforms with the minimum energy possible to fit the data. Because the output of the model is the product of two internal waves, the solution without weight decay is not unique (one wave could always be scaled up if the other were scaled down by the same amount). Introducing the minimum energy constraint gives a unique solution that seems physiologically plausible. The order of presentation of the data set was randomized before each iteration of the training. Thus it is unlikely that resultant waveforms depend on the details of the training. We evaluated the performance of the models on the basis of the residual mse and on the correlation between the models' predictions and the actual waveforms.

RESULTS

The memory-wave hypothesis

We tested our multiplicative model of IT neuronal function on the 21 neurons collected, while the monkey was performing the task with either three or six Walsh patterns. We used these restricted stimulus sets to ensure that there were many repetitions (range 15 to 150, mean 48) of each sample-test combination. The model provided a surprisingly good fit to the data. Figure 3 shows the responses of one neuron (—), along with the fits of the multiplicative model (— —), based on all 36 combinations of the six sample and six test stimuli. Each column shows the responses to one test stimulus as a function of the preceding sample stimulus. Each row shows the responses to different test stimuli that were all preceded by the same sample stimulus. Thus the responses to the matching test stimuli are along the diagonal (boxes). The model fit the shape of the waveforms as well as their overall magnitude.

The model's performance was quantified for each sample and test combination by calculating the Pearson correlation coefficient between the average of the actual response waveforms and the model's predicted waveforms. In this case, the correlation coefficient is a normalized measure of the overall goodness of fit irrespective of the variance of the data set. The model's average correlation coefficient over all stimulus combinations and all neurons was quite good ($r = 0.85 \pm 0.02$ (SE)). We infer from these findings that the responses of IT neurons could be constructed by multiplying two waveforms representing information arising from vision and memory.

As a control, a simpler model was made by assuming that the waveform recalled for comparison was the same as the response to the sample stimulus. This simple model uses the product of the waveforms elicited when each stimulus appeared as the sample. It relies on the assumption that the responses to the sample stimuli are purely sensory. However, the responses to the sample stimuli are likely to be a combination of sensory and memory signals.

The 2 waveforms are then combined, via addition or multiplication, to yield the final output (thick trace). The input layer (squares) encode the Walsh stimuli as vectors in the Walsh transform domain, so that each unit corresponds to a different Walsh pattern. Open circles in the hidden layer represent the combination operation, which may be either addition or multiplication, of the inputs. Each unit in the hidden layer receives a weighted projection from each unit in the input layer. In contrast, each unit in the output layer receives only 2, unweighted, inputs, one from each of the 2 channels at the corresponding point along the trace. Weights of the connections between the input and hidden layers were adjusted until the output matched the actual waveform elicited from a given neuron by the paired stimuli.

The model uses internal waveforms with the minimum energy to fit the data. Because the output of the model is the product of two internal waves, the solution without weight decay is not unique (one wave could always be scaled up if the other were scaled down by the same amount). Introducing the minimum energy constraint gives a unique solution that seems physiologically plausible. The order of presentation of the data set was randomized before each iteration of the training. Thus it is unlikely that resultant waveforms depend on the details of the training. We evaluated the performance of the models on the basis of the residual mse and on the correlation between the models' predictions and the actual waveforms.
This model would predict that the responses would be symmetric, i.e., the response to pattern 2 after pattern 1 would be the same as the response to pattern 1 after pattern 2. In fact, this is not the case (Fig. 3), and this simple model, on average, does not perform as well ($r = 0.47 \pm 0.01$) as the model with no assumptions about the recalled waveform ($r = 0.83$).

The multiplication hypothesis

There are many ways in which the observed waveforms could be constructed from components related to the current and previous stimuli. These components could interact in either a linear or a nonlinear fashion. We were interested in determining whether the nonlinear multiplicative model was necessary to explain the observed data or whether a linear model could explain the data equally well. To answer this question we compared the multiplicative model, the simplest nonlinear model, with an additive model, the simplest linear model.

Figure 4 shows the responses of one neuron (---) along with the fits from the multiplicative (---) and the additive (-----) models. Whereas both models did a reasonable job, the multiplicative model fit the waveforms of the neuronal responses better than the additive model. To compare the performance of the two models, we relied on the mean-squared-error. The mse is an unnormalized measure of the goodness of fit and accounts for the variance of the data set. Thus it is useful for comparing the performance of the two models on the same data set. Analysis of the fits for all 21 neurons revealed that the multiplicative model converged to a lower mse than the additive model for all 21 neurons (sign test, $P < 0.001$). Thus the simplest linear model does not describe the observed data as well as the simplest nonlinear model. This result supports our specific hypothesis that some of the neurons in IT cortex are performing a multiplicative operation on waveforms that might arise from vision and memory.

Test for multiplicative interactions in the data

The multiplicative model has many degrees of freedom because of the large number of weights connecting the input layers to the hidden layers. This raises the possibility that the model could fit any data set whether or not the internal structure of the data reflected an underlying multiplicative mechanism. To control for this possibility, the linkage between the sample-test pairings and the responses...
FIG. 4. Comparison of the fits of the multiplicative and additive models to actual responses. Each panel represents the responses to 1 pair of sample and test stimuli. Graphs show the neuron's mean response waveforms (—) and outputs of the multiplicative (— —) and additive (······) models. Multiplicative model fits the waveform of the response better than the additive model.

was shuffled. That is, each set of responses elicited by a particular sample-test combination was reassigned to another sample-test combination. If the multiplicative model was truly sensitive to consistent features of the actual data, one would expect its performance to be worse on the shuffled data set than on the real data set. On the other hand, if the multiplicative model had too many degrees of freedom, and thus was overly flexible, one would expect its performance on the shuffled data set to equal that of the real data set.

Figure 5 shows the responses of one neuron (—) and the multiplicative model's fit before (— —) and after (······) the data were shuffled. The fit was much worse when the linkage between the sample-test pairings and the responses was disturbed. Analysis of the residual mean-squared-error revealed that the multiplicative model performed less well on the shuffled data than on the real data in all but two of the neurons tested (sign test, $P < 0.001$). Thus, the multiplicative model is not overly flexible, but rather, corresponds to some consistent structural features in the data that are related to combining vision and memory information.

Predicted waveforms from vision and memory

The calculation of a crosscorrelation requires multiplying two waveforms, integrating the product over time, and normalizing the result. Thus the crosscorrelation is a pure measure of the similarity between two waveforms. If IT
neurons were performing the entire crosscorrelation between waveforms arising from vision and memory, one would expect the responses to matching test stimuli to be enhanced. However, as shown in the companion paper, that is not usually the case. Our results are much more consistent with the notion that IT neurons are performing only the multiplication step of this process and not the integration or normalization. Thus the magnitude of the product waveform is a function of both the degree of similarity and the strength of the two constituent waveforms.

Interpreting neuronal function as a multiplicative operation requires some care, because neurons are only active when their net inputs are excitatory. Neurons are silent when their net input is inhibitory, in which case information about the strength and the temporal pattern of the inputs is lost. Thus predictions about multiplicative neurons depend on whether their inputs are all excitatory, all inhibitory, or mixed. The memory-wave hypothesis predicts that the waveforms arising from the memory channel should have the same stimulus-pattern dependencies as those arising from the visual channel, assuming that the inputs from both channels are excitatory. We can test this prediction because the internal waveforms are available for inspection in our model. It is not possible to find a unique set of internal waveforms whose product is a given output waveform, because one waveform could always be multiplied by any other waveform, as long as the other was divided by that same waveform. However, a unique set of waveforms can be obtained if we add the constraint that the internal waveforms have the least possible energy (or, equivalently, minimum vector length). This is achieved in our model by using weight decay during gradient descent (see APPENDIX).

The internal waveforms from vision and memory for each stimulus pattern were compared with the use of the Pearson correlation coefficient. The correlation coefficient, averaged across patterns, was significant for 12 out of 21 neurons ($P < 0.05$). Figure 6 shows the fit of the model (-----) to the responses (-----) of one of these 12 neurons and the internal pattern-dependent waveforms (boxes). Figure 7 shows the similarity of these internal waveforms arising from vision (-----) and memory (-----) when they are superimposed with the same scale. Thus, for many IT neurons, the predicted waveforms arising from vision and memory have similar stimulus-pattern dependencies. These findings support the proposal that some IT neurons receive excitatory temporally modulated inputs from vision and memory that can be used to compare current and previous stimuli.
FIG. 7. A superposition of the predicted internal waveforms arising from vision and from memory (the same neuron as shown in Fig. 6). Predicted waveforms have been normalized to unit length, and those arising from vision and from memory appear to have the same stimulus-dependencies. Correlation coefficients reflect this similarity ($r = 0.66, 0.82, 0.91$).

FIG. 8. The responses, the multiplicative model's fit, and the internal waveforms for a neuron without a consistent increased response for matching stimuli. The 9 graphs at the upper right show the neuron's mean response waveforms (---) and the model's output (-----). The 3 traces in the left column are the model's predictions of the temporally modulated output of the memory system. The 3 traces in the bottom row are the model's predictions of the output of the visual system.
The internal waveforms determined for the remaining nine neurons were not similar. Figure 8 shows the fit of the model (---) to the responses (----) of one of these neurons. Although the fit of the model is good, the internal pattern-dependent waveforms (boxes) are not similar. Figure 9 shows this lack of similarity. One possible explanation of why the multiplicative model could result in a good fit to the data without similar internal waveforms is that one of the input waveforms to the neuron is actually inhibitory. If one waveform is inhibitory, the information needed for the model to accurately predict both waveforms is lost. Nevertheless, it is important to realize that the fit of the model to the data set was very good, even though the internal waveforms were not similar to each other.

**DISCUSSION**

This modeling study supports two new hypotheses of IT neuronal function. The temporal waveforms recorded from IT neurons can be constructed from two sets of waveforms corresponding to current and recalled stimuli. This finding supports the memory-wave hypothesis, i.e., that memory generates stimulus-dependent temporal waveforms similar to those elicited by visual stimuli. The multiplicative model of IT neurons fit the data better than a similar additive model. Furthermore, the fit was better when the responses were linked with the correct sample and test stimuli than when the linkage was shuffled. The results show that the responses of IT neurons are separable into two independent sets of waveforms. This supports the multiplication hypothesis, i.e., that IT neurons multiply the two input waveforms arising from vision and memory. This multiplication might form the basis for a comparison of the waveforms (Fig. 10). These hypotheses provide an explicit definition for the functional role of IT neurons in visual memory tasks.

**Neuronal ensembles and memory waves**

The study of memory has been concerned primarily with the biochemical and biophysical mechanisms of storage or the localization of memory mechanisms to different parts of the brain. Furthermore, theoretical work on memory has failed to explain how the contents of memory can be recalled and used. Our hypothesis suggests that separate neuronal ensembles, one activated by vision to encode the current stimulus, and one activated by memory to encode the previous stimulus, interact in a third ensemble (represented by the IT neurons from which we recorded) that compares the two encodings. Our results suggest that memories, like visual images, are encoded in the temporal modulation of the activity across an ensemble of neurons. Furthermore, the internal memory waveforms appear (in 12/21
neurons) to be capable of representing visual information with the same pattern-dependent temporal modulation as the waveforms elicited from the visual system.

The results obtained in this paper support the general hypothesis that IT neurons combine temporal waveforms arising from vision with temporal waveforms arising from memory. A number of experiments have shown that visual images are encoded by temporally modulated waveforms in various parts of the brain (Cattaneo et al. 1981; Eckhorn and Pöpel 1974, 1975; McErlain et al. 1991; Richmond et al. 1987, 1990). Consequently, it is not unreasonable to think that memories are also manifested as temporally modulated waveforms. Thus, in sequential image comparisons, the memory system might generate a temporally modulated message about the pattern of the previous image. The temporally modulated waveforms related to memory could arise from long-term memory or they could be generated from short-term memory. Our experiments do not allow us to differentiate between the two possibilities. The encoding scheme might be to save the coefficients needed to regenerate the temporal messages as weights on modifiable synapses. The memory could then be recalled by using the stored coefficients to modulate a set of signal generators excited by a trigger signal linked to the onset of the visual stimulus. Elucidating the site and mechanism of such a memory wave system is an exciting challenge for future research.

Multiplicative mechanisms

Our results lead us to the specific hypothesis that IT neurons perform a multiplication of the waveforms arising from vision and memory. The modeling does not prove that a multiplication occurs, but it at least suggests that some nonlinear operation is required to explain the observed waveforms. The ability of the multiplicative model to fit our data is not a reflection of the number of degrees of freedom in our model (as demonstrated by the shuffling procedure, discussed earlier). Any decision making process is inherently nonlinear. Thus, a comparison which ultimately leads to a decision should involve some nonlinear operations, of which multiplication is the simplest. Multiplicative units have been used previously in solving certain classes of problems, such as those found in modeling attentional processes (Durbin and Rumelhart 1989). Furthermore, multiplicative units are not physiologically implausible. For example, voltage sensitive channels, such as NMDA receptors, can introduce a nonlinear operation at the membrane level that is quite similar to multiplication (Collingridge and Bliss 1987). In addition, it has been found that NMDA receptors provide a multiplicative gain control mechanism in cat visual cortex (Fox et al. 1990).

Integrative decision process

While the waveforms from vision and memory are being multiplied, it would be a relatively straight-forward matter to integrate their product over time. This integration would eliminate pattern-dependent information, leaving only the similarity measure. Once a threshold is reached, a decision could be made regarding the similarity of the two waveforms. This decision would then be reflected in the monkey's motor response. At the present time, we have no evidence of where this proposed integration might occur.

Computational theory of cognitive tasks

There are no established quantitative theories of IT neuronal function. Recently, Miller et al. (1991) proposed that "IT neurons may be acting as adaptive mnemonic filters that seek to preferentially pass information about new, unexpected, or not recently seen stimuli." However, our results showed that the mean firing rate averaged across all of our neurons was the same in sample, match, and nonmatch conditions. Furthermore, we found that the amount of information transmitted about stimulus pattern is the same in sample, match, and nonmatch conditions (Eskandar et al. 1992). Thus we can only conclude that the adaptive mnemonic filter hypothesis is, at best, limited to only some neurons under some circumstances. A more general hypothesis of IT neuronal function must account for the wide range of neuronal responses observed under different conditions. This paper shows that our hypothesis explains the temporally modulated responses, including the changes in mean firing rate, in all neurons, whether the matching responses were the same, larger or smaller. Hence, our hypothesis includes the effects that gave rise to the adaptive filter hypothesis, and also accounts for our other data, which could not be explained by the adaptive filter hypothesis.

We have proposed two new hypotheses of neuronal function in visual memory. The memory-wave hypothesis states that the memory system generates a temporal waveform representing the recalled information. The multiplication hypothesis states that neurons in area TE of IT cortex participate in comparing current and recalled stimuli by multiplying the waveforms arising from vision and memory. These hypotheses show that a computational theory of a cognitive task requiring visual memory can be based on the temporal modulation present throughout the visual system.

APPENDIX

The models used to decompose the recorded neuronal responses into hypothetical visual and memory waveforms had the structure of a modified three-layer artificial neural network. There were variable weights between the input and hidden layers, and the output layer had an adjustable bias. The models were trained by adjusting these weights and biases to minimize an error function. The error function, $E$, was defined as the sum of two terms. The first term, $E_0$, was the expected value of the mean squared error between the model predictions and the actual responses:

$$E_0 = \frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{L} (y_{jk}^{\text{model}} - d_{jk})^2$$

where there are $L$ data points (64 in this case) and $M$ trials (usually several thousand). The second error term, $E_1$, was defined as the total energy in the two internal waveforms, before passing through the compressive nonlinearity (which was usually not a factor in these tests). Because the weights are represented as vectors, $E_1$ is just the sum of the squares of the two vector elements:
where $w_{ij}$ is the set of internal weights from channel 1, and $w_{2j}$ is the set of internal weights from channel 2. The relative importance of reducing the residuals and minimizing the weights is controlled by $\lambda$ (0.001 in this study).

For the multiplicative model, the output at each point in layer 3 is defined as:

$$y_i^{(3)} = y_{1i}^{(2)} \times y_{2i}^{(2)} + b_i \quad (A3)$$

where $y_{1i}^{(2)}$ is the input from channel 1, $y_{2i}^{(2)}$ is the input from channel 2 of the hidden layer, and $b_i$ is the bias of that unit. For the additive model, the output is just:

$$y_i^{(3)} = y_{1i}^{(2)} + y_{2i}^{(2)} + b_i \quad (A4)$$

The model’s parameters were adjusted to minimize the error function. To derive the update equations of the weights for gradient descent, we obtained the gradient in the hidden layer from the gradient in the output layer by the chain rule (Rumelhart et al. 1986). The units in the hidden layer have outputs:

$$y_i^{(2)} = P(x_i^{(2)}) \quad (A5)$$

where

$$P(x) = \frac{1}{1 + \exp(-x)} \quad (A6)$$

and $x_i^{(2)}$ is the total input to that unit, given by:

$$x_i^{(2)} = \sum_j w_{ij} x_j^{(1)} \quad (A7)$$

Now we can define the change in the biases by the learning rule:

$$\Delta b_i = \eta \frac{\partial E}{\partial b_i} \quad (A8)$$

but from the chain rule:

$$\frac{\partial E}{\partial b_i} = \frac{\partial E_0}{\partial y_i^{(2)}} \frac{\partial y_i^{(2)}}{\partial b_i} \quad (A9)$$

where from Eq. (A1):

$$\frac{\partial E_0}{\partial y_i^{(2)}} = (y_i^{(3)} - d_i) \quad (A10)$$

and from Eq. (A3) or (A4):

$$\frac{\partial y_i^{(3)}}{\partial b_i} = 1. \quad (A11)$$

To get the change in the weights in the inner layer, we can use the learning rule:

$$\Delta w_{ij} = \eta \frac{\partial E}{\partial w_{ij}} \quad (A12)$$

Applying the chain rule:

$$\frac{\partial E_0}{\partial w_{ij}} = \frac{\partial E_0}{\partial y_i^{(2)}} \frac{\partial y_i^{(2)}}{\partial x_j^{(1)}} \frac{\partial x_j^{(1)}}{\partial w_{ij}} \quad (A13)$$

where $i$ is 1 or 2, and $j$ ranges over the length of the vector.

For the multiplicative model (Eq. A3) there are two different partial derivatives, one for each channel:

$$\frac{\partial y_i^{(3)}}{\partial y_j^{(2)}} = \frac{\partial y_i^{(3)}}{\partial x_j^{(2)}} = \lambda w_{ij} \quad (A14)$$

For the additive model (Eq. A4) the partial derivatives in both channels are the same:

$$\Delta w_{ij} = \frac{\partial E_0}{\partial w_{ij}} \quad (A15)$$

and from Eq. (A5):

$$\frac{\partial y_i^{(3)}}{\partial x_j^{(2)}} = \lambda w_{ij} \quad (A16)$$

and from Eq. (A7):

$$\frac{\partial x_j^{(2)}}{\partial w_{ij}} = y_i^{(3)} \quad (A17)$$

The authors thank Dr. John Hertz, of NORDITA, for the invaluable discussions on the use of connectionist networks in neural modeling. E. N. Eskander was supported by the Howard Hughes Medical Institute NIH Scholars program.

Received 12 November 1991; accepted 29 May 1992.

REFERENCES


Fon, M., Daw, N., and Sato, H. The effect of varying stimulus intensity...