What is a moment? 'Cortical' sensory integration over a brief interval (I) 7 October 2000

J. J. Hopfield
Department of Molecular Biology
Princeton University, Princeton NJ 08544-1014

Carlos D. Brody
Center for Neural Science, New York University
4 Washington Place Room 809 NY, NY 10003

Abstract

Recognition of complex temporal sequences is a general sensory problem that requires integration of information over time. We describe a very simple 'organism' that uses novel neural computational principles to perform this task, exemplified here by recognition of spoken monosyllables. The 'organism' is a network of very simple neurons and synapses; the experiments are simulations. The network's recognition capabilities are robust to variations across speakers, simple masking noises, and large variations in system parameters. The network principles underlying recognition of short temporal sequences are applied here to speech, but similar ideas can be applied to aspects of vision, touch, and olfaction. In this paper, we describe only properties of the system that could be measured if it were a real biological organism. We delay publication of the principles behind the network's operation as an intellectual challenge: the novel essential principles of operation can be deduced based on the experimental results presented here alone. An interactive web site (http://neuron.princeton.edu/~moment) is available to allow readers to design and carry out their own experiments on the 'organism'.

Introduction

How does a brain integrate sensory information that occurs over a time on the scale of ~0.5 seconds, transforming the constantly changing world of stimuli into percepts of a 'moment' of time? This is a general problem essential to our representation of the world. In audition, the perception of phonemes, syllables, or species calls are examples of such integration; in the somatosensory system, the feeling of textures involves such integration; in the visual system, object segregation from motion and structure from motion require short-time integration; in the olfactory system, sensing odors during a sniff involves temporal integration. Linking together recently-occurred information into an entity present 'now' is a fundamental part of how the percept of a present 'moment' is constructed; a key issue in this regard is how such integration over time can be carried out using neural hardware.

We describe here a simple and very biologically plausible network of spiking neurons that recognizes short complex temporal patterns. In so doing, the network links together information spread over time. The network was designed using novel computational principles. It is capable of broad generalization from a single example and is robust to noise. These capabilities are demonstrated here by considering the real-world problem of recognizing a brief complex sound (a monosyllable; see Figure. 1). We chose this representative but specific task because it is a natural capability of our auditory systems. The task is well-defined and conceptually easy to describe, and real-world data is available to exemplify the important problem of natural variability and noise. The object here is to understand how high selectivity for spatiotemporal patterns can be obtained in a biological system; the performance of this simple system as a word recognizer is of course far worse than digital computer based commercial systems, but the comparison is not relevant.

In this paper, we describe the network by presenting only observations and experiments that would have been performed on the network if it were a real biological organism. As with a real organism, we do not explicitly describe in this paper the principles underlying the network's operation, but merely the experimental facts that one can record about it; the principles of operation must be deduced. In a few months, we will present in a second paper a full and explicit description of the principle behind the design and performance of the system.

We have chosen this unusual mode of presentation based on our own experience with the system. We were surprised to find that the information described in this paper was sufficient to deduce the principles on which the system works. Had this been a real biological system, we ourselves would have been inclined to believe instead that the secret of the specificity must lie in additional cell types, cellular biophysical complexity, or other as-yet unmeasured fundamental properties; we would have glossed over subtleties that actually clearly indicate how the system works, and we would never have found the interesting principle on which the network computation is based.

How often have we been guilty of similar behavior in looking at data from neurobiology? Probably often, from lack of practice in interpreting data instantiating a new principle in an unusual fashion. How often have others been guilty of similar behavior? We cannot know; probably less than ourselves. However, feeling that some in the community might appreciate an opportunity to interpret the behavior of a system that is guaranteed to have no hidden components, we have chosen to present in this paper only conventional 'experimental information'—conventional information that we know is sufficient to derive the underlying computational principle and to understand how the system computes. Someone who simply wants to be told the principle will find that information in the second paper, to be published shortly after this one.

We will begin by describing the network's complex pattern recognition behavior, and the firing patterns of those neurons that are correlated with this behavior. We will then turn to the full network, and describe its neuroanatomy (cell types, synapse types, connectivity pattern), physiological properties in response to acoustic stimuli, and single-cell properties as observed *in vitro*. We will write as though the network were a real organism, as far as the experimental measurements are concerned.

Behavior and electrophysiological correlates

In the particular network described here, a novel principle of computation that enables the network to recognize the spoken monosyllable "one," as well as 9 other different patterns, has been instantiated through a particular choice of network parameters. The method to determine the parameters that enable recognition of each pattern is non-iterative, and requires information based on only a single example of the pattern to be recognized. Although the method is straightforward, it is not the focus of our study and we will not describe it further, only noting that we believe the

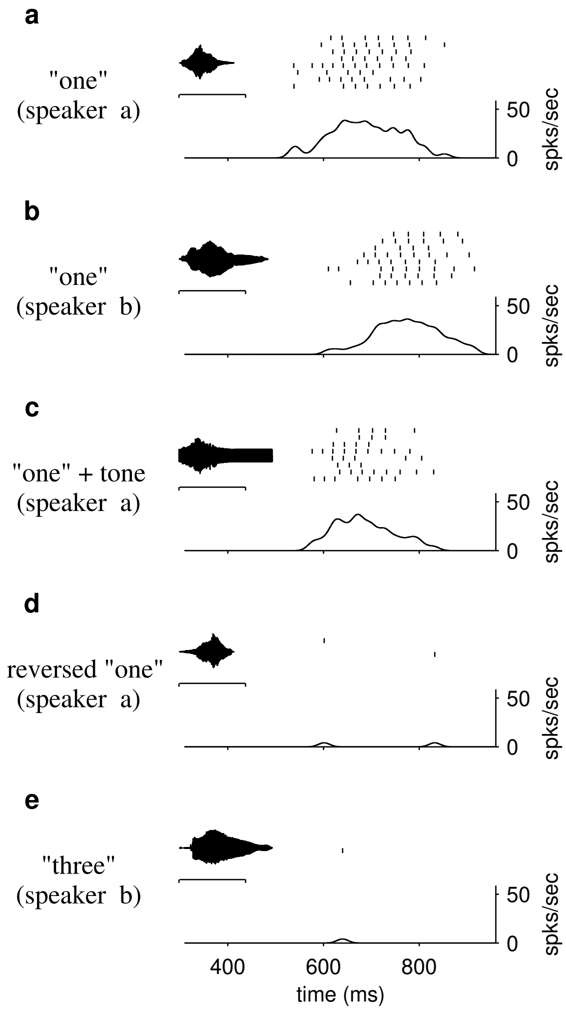

Fig 1 Extracellularly recorded responses of a single γ-type neuron to five different acoustic waveforms. A noisy membrane current was added to every neuron in the simulation of the neuronal mathematics for the 'organism', to simulate the noise due to other inputs that would always be present in a real biological system. Before the experiment, the network parameters were set using only a single exemplar of the word 'one' spoken by speaker (a), plus single examples of 9 other different patterns (each recognized by one of 9 other γ neurons, not shown here.). a) Spike rasters, aligned in time to the start of the acoustic waveform shown in the inset, in response to 8 different trials using an utterance of the word 'one' from speaker (a) (not the training exemplar). Below the rasters is their corresponding peristimulus time histogram (PSTH), smoothed by a Gaussian with a standard deviation of 12 msec. The γ cell begins spiking near the end of the word. Tick marks in the inset correspond to 0 and 500 ms. b) Same format as panel A, for an utterance of the word 'one' from a different speaker (b). c) Same format as panel A for a 'one' spoken by speaker b in the presence of a loud tone at 800 Hz.The waveforms are markedly different in a),b) and c) yet the γ cell responds to all. d) Same format and utterance as in panel a), but the acoustic waveform has been reversed in time. e) Same format as panel a), for an utterance by speaker (b) of the word 'three'. Few or no spikes occurred in response to the waveforms of panels d) and e). Other, similar-sounding words (for example, 'wonder') occasionally cause the cell to fire as well, indicating that these output cells are not completely specific, but merely encode utterances quite sparsely.

parameters could also be set by biologically plausible synaptic learning rules. ‘Recognition’ of each of the 10 patterns is signaled by the firing of a corresponding pattern-selective neuron. We have labeled such neurons ‘γ’ neurons (see Anatomy section below), and will focus on the behavior of the particular γ neuron that is selective for the word ‘one’. The neuron fires in response to this word, whether it is spoken rapidly or slowly, or when spoken by a variety of speakers. When a loud sound at 800 Hz is played simultaneously with the word ‘one’, the network’s ability to recognize the word is only slightly degraded. In contrast, the γ neuron does not respond to ‘one’ played backward or to most monosyllabic utterances, although on occasion it does respond to words which are similar to ‘one’. In short, the system contends with the kind of natural variations and context with which humans can contend, and has a good ability to reject simple masking sounds.

Data from one γ neuron is shown in Figure 1, and illustrates the selectivity of the neuron’s response to simple sound stimuli. Figs. 1a-c illustrate the response to the word ‘one,’ spoken by two different speakers and in two very different acoustic contexts. The neuron responds robustly in all three cases. In contrast, as illustrated in Figs. 1d-e, the neuron responds weakly or not at all to other utterances, despite their superficial similarity to the word “one.” γ neurons do not respond to pure sinewave tones (data not shown).

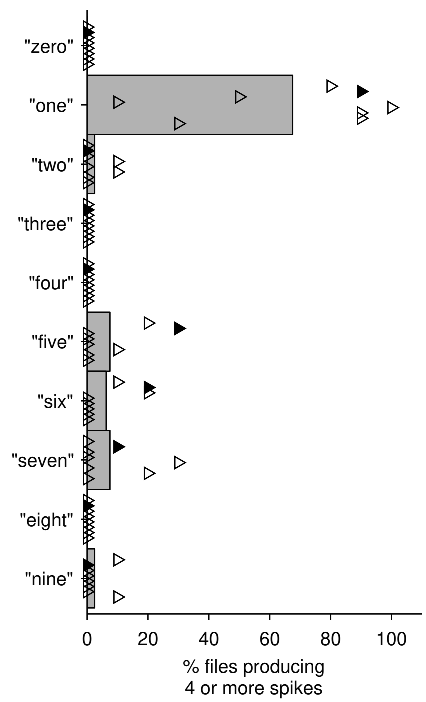

Fig 2 Summary of responses of a single γ cell to ten spoken digits, ‘zero’ through ‘nine.’ (Speech data taken from the TI46 database, available from NIST.) Each digit was spoken ten times by eight different female speakers while the responses of the γ cell were recorded. For the purpose of evaluating the cell’s selectivity, each trial was classified as ‘responding’ if the γ cell fired 4 or more spikes, and as ‘not responding’ otherwise. Triangles indicate averages over different utterances by individual speakers, while the gray bars indicate data averaged over all utterances of all speakers. For 5 of the 8 speakers, the cell’s response is highly selective for the word ‘one.’ The filled symbol indicates the speaker from which the single training utterance was taken.

Figure 2 illustrates the result of stimulating the system with a variety of spoken digits. For most speakers, the neuron was highly selective for the word 'one'. Most of the failures to respond to "one" were on utterances of three speakers on whom the system had not been trained (lower three triangle symbols in column marked "one" in Fig. 2). This is perhaps not surprising in view of the fact that the parameters for this pattern had been set based on a single example from another speaker. More surprising is the fact that the γ neuron generalized from a single utterance of the training speaker to most utterances of 4 other speakers.

As we will show, the system's complex word-recognition calculation, which in this case involves integrating information spread out over ~0.5 sec, is carried out by cells that have remarkably simple biophysical and physiological properties. The network's neurons can be well-described as a straightforward collection of classical integrate-and-fire neurons with elementary synaptic connections between them.

Neuroanatomy

We describe the architecture of the network as if it were arranged in a biological-like layout. γ neurons are found grouped into the superficial layers of what we dub here 'area W'. Auditory information reaches area W via another area, 'area A', which may be thought of as a cortical sensory area. Neurons in area A are frequency-tuned (see "Electrophysiology" section below), and are arranged in groups having similar preferred frequencies, that is, frequency is tonotopically mapped. Output neurons of area A project to what we have called layer '4' of area W. Word-selectivity arises in area W, and we will therefore focus our anatomical description on area W.

The axons of area A output cells arborize narrowly in layer 4 of area W, and preserve the tonotopic mapping found in area A. Layer 4 of area W contains two types of cells, both of which receive direct excitatory synaptic input from area A afferents. α type cells are excitatory, and β type cells are inhibitory. Both of these types of layer 4 cells are found in similar numbers. All cells are electrically compact.

The axons of both α and β neurons arborize widely within area W, each making a total of approximately 75-200 synaptic connections with other neurons, across all tonotopic frequency groups. Approximately half of the connections from each cell are onto α cells, the other half are onto β cells. Axons of α and β cells also arborize in layers 2 and 3, where they contact γ cells.

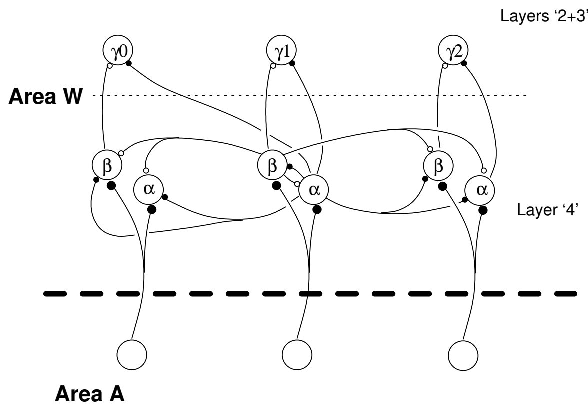

Fig 3 Schematic neuroanatomy for area W and its input. The thick dashed line separates area A from area W; the thin dotted line separates layers '2+3' from layer '4' in area W. Small filled circles indicate excitatory connections, while small open circles indicate inhibitory connections. The connections of a typical α cell and a typical β cell, both shown in the center, are sketched. In the simulations, area W is small, containing 325 neurons of each α and β type, and a given cell makes synapses on 15-30% of these cells. Our simulation contains 10 different γ cells, each selective for a different temporal pattern. Each γ cell receives inputs from 30-80 cells of each type α,β.

There are about 3% as many γ cells as there are α or β cells. Each γ cell receives approximately 30-80 synapses from cells of type α and of type β; these inputs are drawn from cells in all frequency groups. γ cells are the output cells of this system: their axons project to other cortical areas, where they make excitatory synapses. They do not feed back to α or β cells.

Distances within area W are short, and the diameters of axons of all cell types are large. Thus, propagation delays within area W appear to be unimportant. Latencies from area A to area W are the same for all cells.

Electrophysiology *in vivo*

Area A

As described above and shown in Fig. 3, the projection neurons of area A provide the input to area W. We continue our description in the language of neurobiological experiments, but note that the neural interactions important for word selectivity are found in area W; the mechanisms that give rise to the properties of area A neurons are not of relevance. The detailed (non-biological) source code for area A in the simulations can be found on the main web site associated with this paper.

The properties of area A neurons can be summarized by saying that (a) area A neurons are frequency tuned; and (b) the neurons respond to transient changes in acoustic signals with a train of action potentials of slowly decaying firing rate. Cells in area A responded transiently to three types of "features": onsets (~ 35% of cells), offsets (~ 35%), and peaks (~ 30%) of power in modulated sinewave tones. The cells exhibited no tonic response to continuing steady sounds of any frequency. Every response produced a slowly decaying train of action potentials after initiation. (See Figure 4a.) Different cells had different response decay rates: Figure 4b illustrates two 'on' cells with different decay times. Over the population of recorded neurons, a wide variety of decay.

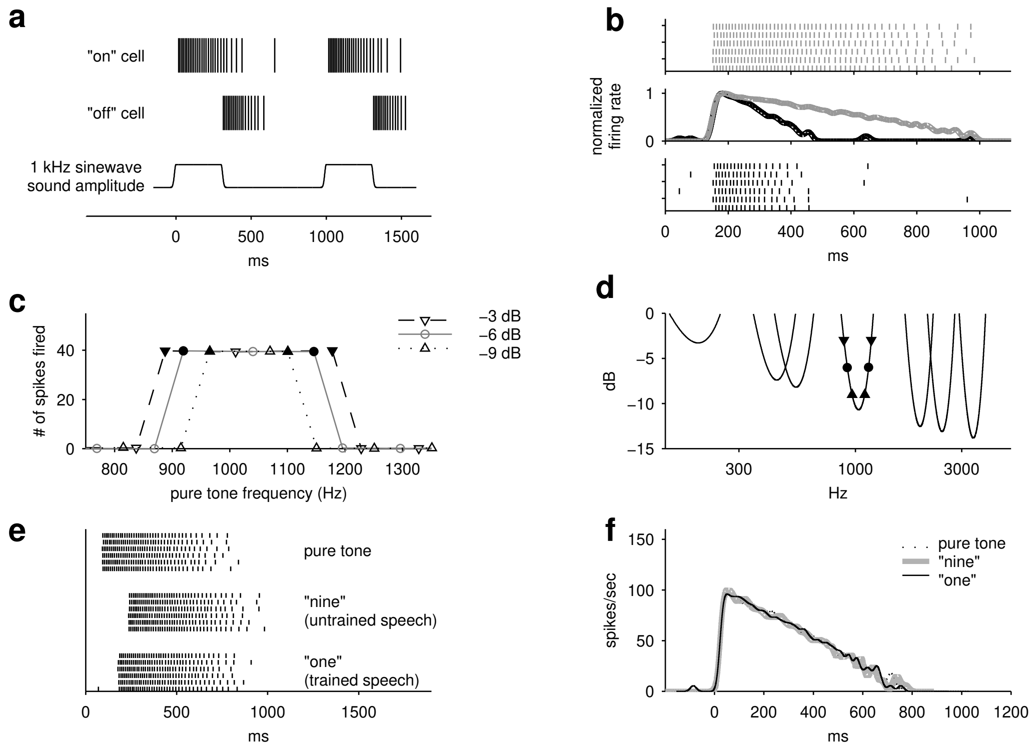

Fig. 4 a) Spike rasters for a typical 'on' cell and a typical 'off' cell in response to two pure sinewave tone stimuli, as indicated at the bottom of the panel. The beginning and end of each tone are slightly smoothed as shown to minimize the generation of spurious frequencies by the sharp transient. b) Responses of two different 'onset' cells to six different trials of a pure tone onset. One cell is shown in gray, the other in black (top and bottom of panel). The center panel shows PSTHs of the responses of the two cells. c) The number of spikes generated in response to 'step' sinewave inputs (as shown in panel (a)) as a function of sinewave frequency, plotted for three different sinwave amplitudes. Signal power is measured in decibels relative to an arbitrary reference power. As long as the frequency is within a range that depends on signal power (larger range for larger signal powers), the number of spikes generated varies little. Filled symbols indicate the boundary between presence and absence of a robust spiking response. d) Parabolic fits to measurements of threshold power vs frequency, for 7 different onset cells. Each parabola represents a single cell. Filled symbols correspond to filled symbols in panel (c). e) The response of an 'onset' cell to three different stimuli, a pure tone onset, the word 'one', and the word 'nine'. f) Histograms of the responses of panel e) time-shifted into best alignment. When shifted into alignment, there is no apparent difference between these histograms, or between the spike rasters of the three utterances.

times were found, ranging uniformly from 0.3 sec to 1.1 sec. Once a cell initiated a response, subsequent features in the sound stimulus that occurred during the decay had no effect on the spike train. After the end of the decay, newly occurring features can reinitiate the response. (Nevertheless, for simplicity, the particular simulations shown on the web site associated with this paper were restricted to use only the first feature detected, without the possibility of reinitiation.)

Cells in area A were found to be frequency-tuned. Fig. 4c shows the response of a typical 'onset' cell as the power and frequency are varied. The cell responds to a small range of frequencies, the width of which grows with the power of the signal. All cells were found to be frequency-tuned in this sense. Within each cell's range of response-producing frequencies, the cells displayed an almost 'all-or-none' response: provided the signal intensity was above a minimum threshold, each cell fired almost the same number of spikes regardless of the frequency or intensity of the signal. Figure 4d illustrates the frequency tuning of 7 different cells over a range of signal powers. Different signals that were successful in driving area A neurons did not seem to produce significantly different responses. Figure 4e-f shows that the stereotyped response of a typical 'onset' cell in area A was essentially identical for three very different acoustic stimuli that drove it.

For each 'flavor' of area A cell (onset, peak, or offset), different cells, with preferred frequencies spanning the entire frequency spectrum, were found. For each flavor and for each preferred frequency, cells with a broad range of decay rates were found.

Area W

We now turn to the electrophysiology of neurons in area W, where word-selectivity arises. As in area A, cells of both type α and β in layer 4 of area W are arranged in groups with similar preferred frequencies. The responses of both α and β cells were found to be similar to the output cells of area A which drive them: the three types 'onset', 'offset', and 'peak' cells were all found in layer 4 of area W. As in area A, for each 'flavor' of α or β neuron (onset, peak, or offset), different cells, with preferred frequencies spanning the entire frequency spectrum, were found. For each flavor and for each preferred frequency, cells with a broad range of decay rates were found. The responses of one 'onset' and one 'offset' cell are illustrated in Fig. 5a. Amplitude steps of pure sinewave tones were used to drive two 'onset' cells with different decay rates, illustrated in Fig 5b.

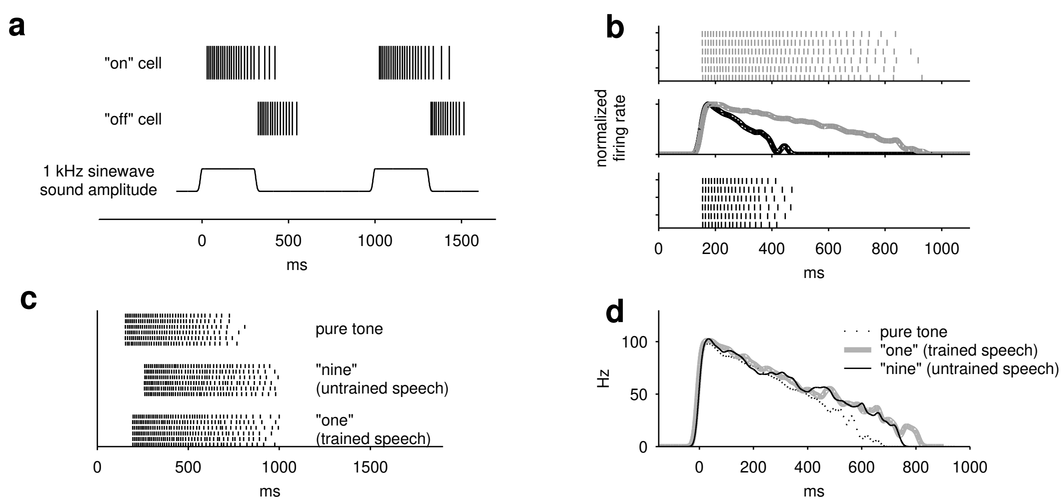

Fig. 5 Responses of layer 4 area W cells. a) The spike rasters for a typical 'on' cell and a typical 'off' cell in response to sine wave pulses; format as in Fig4a. b) Responses of two different onset cells to six different trials using the same pure tone onset; format as in Fig 4b. c) The response of an 'onset' cell to three different stimuli, a pure tone step, the word 'one', and the word 'nine'; format as in Fig 4e. d) Histograms of the responses of panel e) shifted into a common response onset time; format as in Fig 4f.

In sum, when studied with pure tones, the decay rates and frequency tuning properties of α and β cells were very similar to those described for area A (see Figure 4). In contrast, small but reliable differences were found when the cells were stimulated with speech signals. Fig. 5c illustrates the responses of an 'onset' cell to three different stimuli, a pure tone step, an utterance of the word 'nine' (on which the animal had not been trained), and an utterance of the word 'one' (a word on which the animal had been trained). Figure 5d shows the PSTHs of these responses, aligned to a common response onset time. While in area A the responses to pure tones and speech are indistinguishable from each other, in area W the PSTHs of the responses to speech are subtly but consistently different to the PSTH of the response to the pure tone. After approximately 400 ms, the response to speech signals is consistently stronger and more persistent than the response to pure tone steps. Thus, layer 4 of area W is the first level of the pathway leading to the word-selective γ cells of layers 2+3 that shows a response component that is specific to speech. We do not know the precise role that this late sustained component may play in word-selectivity. Firing patterns and response properties of α and β cells seem essentially identical.

The characteristics of the layer 2-3 "one"-selective γ cell shown in Fig 1 have already been described. The simulation contains 9 additional γ cells, each 'tuned' to detect a different pattern composed of a randomly chosen arrangement of onsets, peaks, and offsets. The selectivity properties of each of these γ cells, with respect to their target pattern and variants around it, are similar to that of the "one"-selective γ cell, and we do not describe them further here. When γ cells respond, they generally do so with a pattern containing 4-8 spikes with a typical 'frequency' of 30-60 hz.

Intracellular recording in a slice preparation

Finally, we turn to *in vitro* studies of the biophysical properties of the α, β, γ neuron types in area W. The three cell types are qualitatively similar, and appear to be well-described by simple integrate-and-fire cell models.

Synapse properties were studied using conventional two electrode methods. Excitatory postsynaptic currents (EPSCs, illustrated in Fig. 6a) have an extremely fast rise time and decay exponentially with a time constant of 2 ms. Inhibitory postsynaptic currents (IPSCs, illustrated in Fig. 6b), in contrast, have a slower rise time. IPSC waveforms were well-fit by alpha-functions (fits not shown), with a peak amplitude time of 6 ms. The recordings shown in Figs. 6a-b were made with cells held at –65 mV, but waveform amplitudes and time constants changed little when the holding potential was varied within the range –75 mV to –55 mV. Paired-pulse experiments (data not shown) have demonstrated that both excitatory and inhibitory synapses in area W neither adapt nor facilitate, and that synaptic currents due to closely timed action potentials add linearly. The EPSPs and IPSPs obtained in the same cells as shown in Figs. 6a-b when the voltage clamp was removed are shown in Figs. 6c-d.

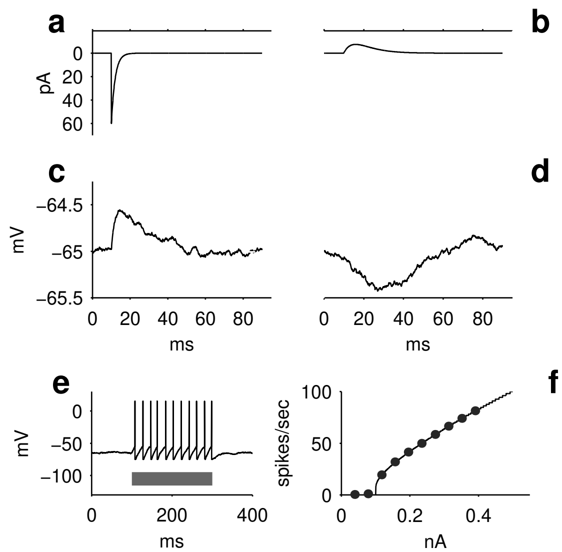

Fig. 6 Whole-cell recordings from α and β cells in layer 4. Panels a-d: A minimal stimulation protocol was used to observe synaptic responses due to the activation of a single axon afferent to the recorded cell. a) Excitatory postsynaptic current measured in a β cell under voltage clamp conditions. b) Inhibitory postsynaptic current measured in an α cell. c) EPSP, measured in the same cell as in (a). Resting state here corresponds to the cell's resting membrane potential, -65 mV. (Since noise is present in all real biological systems, here and in all other simulations, independent white Gaussian noise with standard deviation 0.2 mV was added to the neuron's membrane potential at each 0.1 ms timestep.) The trace shown is the average of 1000 repeats. d) IPSP, measured in the same cell as in (b). e) Spiking response to an above-threshold current step, showing no spike-frequency adaptation. Grey bar indicates the time during which current was injected. f) Firing rate of an α cell as a function of input current. Points are the experimental measurements, and the solid line is a calculated fit to these points, based on a leaky integrate-and-fire model of the cell.

α and β cells were found to be electrotonically compact, with membrane time constants of approximately 20 ms. We applied a series of constant current steps of different amplitude to these cells. One such application is illustrated in Fig. 6e, and the result of the entire series of steps is summarized by the data points shown in Fig. 6f. By all studies we have made, both α and β cells have properties which can be duplicated by leaky integrate-and-fire neurons with a short absolute refractory time period. The solid line in Fig. 6f is the result of fitting such a model to the data points shown in the same panel. The parameters of the fit were: absolute refractory time period 2 ms, membrane time constant 20 ms, resting potential –65 mV, firing threshold –55 mV, reset potential after spiking –75 mV, membrane capacitance 250 pF. Though *in vivo* firing rates greater than 150 hz are seldom seen, when driven by steady currents the maximum firing rate of all three cell types is around 500 spikes/sec.

The γ cells of layer 2+3 cells are qualitatively similar to α,β cells in every way, but quantiatively γ cells have a smaller membrane resistance, and a shorter membrane time constant of 6 ms. IPSC's and EPSC's seen in γ cells have time courses very similar to those seen in α,β cells (Fig.6) but the typical peak currents are about three times as large.

Conclusion

This system carries out a difficult computation in a manner that results in significant robustness to variability and noise. It recognizes whole sound sequences in a way that is not sensitive to the kinds of variability that are present in natural vocalizations—variations in voice quality, in speed of speaking a syllable, and in sound intensity. The computation integrates a short epoch of the past into a 'present' decision, in this case about the category to which a recent sound belongs. Despite the complexity and robustness of the computation, the elements that compose the system, and the inputs to it, are remarkably simple. Most importantly, they are similar to the elements found in real neurobiology: our goal is to understand how neurobiology might integrate and recognize spatiotemporal patterns. The biophysics of individual neurons and synapses is that of classical integrate-and-fire neurons with non-adapting synapses. The projection neurons of area A respond to stimuli in a fashion not unlike some responses available in processing regions of a variety of sensory modalities. In short, given apparently ordinary input and computing elements, the system robustly carries out the complex task of recognizing spoken words.

The algorithm effectively carried out by this network of simple, biologically plausible neurons is a novel one, applicable to a wide variety of inputs and situations, whose principles will be described in the subsequent paper. Here we only remark that consideration of the details given will convince a reader that we have not clothed a backprop-trained network in biologically-plausible camouflage, and that the network is using neurons collectively and not as logic elements.

Any neurobiological computation should be robust to cellular variations. For example, high accuracy in individual synapse properties is biologically unrealistic, and a computational neurobiologist will not take seriously schemes requiring great accuracy of individual synapses. The present system itself is 'biological' in this regard--simultaneously varying each synaptic connection strength between neurons in area W by a different random factor of +/-50% has no appreciable effect on the response of the α, β, or γ cells.

The article with an explicit description of the principles of operation of the system will be presented in a few months. Our surprise in finding that the novel network principles could be fully deduced, based on the straightforward experimental results presented here, leads us to ask whether long chains of logical deduction similar to the one appropriate in this case could be usefully applied in neurobiology. We do not claim to know the answer to this question, but we believe it is useful to raise it.

We assure the reader that no additional features of neurobiology are required beyond those herein described either explicitly or implicitly. However, some readers may wish to learn more details, or may wish to carry out experiments of their own design. We have constructed an interactive web site, http://neuron.princeton.edu/~moment, where this can be done. The web site contains speech files with numerous examples of spoken utterances, sinewave pulses, and the corresponding recordings that would be available from single-electrode extracellular studies of the output cells of area A cells, and the α, β, and γ cells of area W. The sound files can be heard, and the sound files and spike rasters can be downloaded. A user can also upload new sound files to the website and study the 'electrophysiology' of responses to those files.

The second paper will contain references to the relevant artificial and biological neural literature. The challenge presented in this paper has involved describing a computational network as if it were a biological one. In this spirit, references to relevant material that led the creators of the network to the principles underlying it are not appropriate, and we have chosen to include none in this paper.

The research at Princeton University was supported in part by NSF grant ECS98-73463 and at New York University by a postdoctoral fellowship to CDB from the Sloan Foundation.