Response Latency of Macaque Area MT/V5 Neurons and Its Relationship to Stimulus Parameters

S. E. RAIGUEL, D.-K. XIAO, V. L. MARCAR, AND G. A. ORBAN

Raiguel, S. E., D.-K. Xiao, V. L. Marcar, and G. A. Orban. Response latency of macaque area MT/V5 neurons and its relationship to stimulus parameters. J. Neurophysiol. 82: 1944–1956, 1999. A total of 310 MT/V5 single cells were tested in anesthetized, paralyzed macaque monkeys with moving random-dot stimuli. At optimum stimulus parameters, latencies ranged from 35 to 325 ms with a mean of 87 ± 45 (SD) ms. By examining the relationship between latency and response levels, stimulus parameters, and stimulus selectivities, we attempted to isolate the contributions of these factors to latency and to identify delays representing intervening synapses (circuitry) and signal processing (flow of information through that circuitry).

First, the relationship between stimulus parameters and latency was investigated by varying stimulus speed and direction for individual cells. Resulting changes in latencies were explainable in terms of response levels corresponding to how closely the actual stimulus matched the preferred stimulus of the cell. Second, the relationship between stimulus selectivity and latency across the population of cells was examined using the optimum speed and direction of each neuron. A weak tendency for cells tuned for slow speeds to have longer latencies was explainable by lower response rates among slower-tuned neurons. In contrast, sharper direction tuning was significantly associated with short latencies even after taking response rate into account, \( P = 0.002 \), ANCOVA. Accordingly, even the first 10 ms of the population response fully demonstrates direction tuning. A third study, which examined the relationship between antagonistic surrounds and latency, revealed a significant association between the strength of the surround and the latency that was independent of response levels (\( P < 0.002 \), ANCOVA). Neurons having strong surrounds exhibited latencies averaging 20 ms longer than those with little or no surround influence, suggesting that neurons with surrounds represent a later stage in processing with one or more intervening synapses. The laminar distribution of latencies closely followed the average surround antagonism in each layer, increasing with distance from input layer IV but precisely mirroring response levels, which were highest near the input layer and gradually decreased with distance from input layer IV. Layer II proved the exception with unexpectedly shorter latencies (\( P < 0.02 \), ANOVA) yet showing only modest response levels. The short latency and lack of strong direction tuning in layer II is consistent with input from the superior colliculus. Finally, experiments with static stimuli showed that latency does not vary with response rate for such stimuli, suggesting a fundamentally different mode of processing than that for a moving stimulus.

INTRODUCTION

Although the primate brain analyzes incoming visual information with surprising rapidity, there is nonetheless a finite delay, or latency, between the time that a visual stimulus appears on the retina and the time that a neuron in the visual system begins to spike in response to that stimulus. This delay arises from a number of factors, including photoreceptor transduction, neural conduction time, synaptic delay, and spike integration time, and tends to increase in higher cortical areas with each successive stage adding its processing time before passing the signal along to the next, higher area (Nowak et al. 1995; Raiguel et al. 1989; Schmolesky et al. 1998; Vogels and Orban 1994). Although it is perhaps not surprising that latencies can vary considerably among individual neurons within a visual area given the number of paths by which information can arrive, the breadth of that range can be remarkably wide (for review, see Nowak and Bullier 1997), far exceeding the average differences between hierarchically adjacent visual areas. Area MT/V5 is no exception, and the range of latencies measured for individual neurons there easily exceeds 100 ms (Maunsell 1987; Raiguel et al. 1989).

The range of latencies in MT/V5 is perhaps more striking insofar as this area receives a restricted neural input almost entirely magnocellular in origin (Maunsell et al. 1990) and comprises but a single retinotopic map composed of a relatively homogeneous population of neurons giving directionally selective responses to translational motion. (Desimone and Ungerleider 1986; Maunsell and Van Essen 1983a). Other factors must therefore account for the wide range of latencies observed in this cortical area: first, although the input may be predominantly magnocellular, that contribution can arise from primary visual cortex by either direct afferents from V1 or indirectly through V2 (Ship and Zeki 1989a,b). Other pathways bypass striate cortex completely (ffytche et al. 1995), either passing through the superior colliculus and pulvinar (Standage and Benevento 1983; Ungerleider et al. 1984) or using direct connections with the LGN (Fries 1981; Yuki and Iwai 1981). Reciprocal connections with areas V3 and V4 also have been described (Maunsell and Van Essen 1983b). Second, the axons within a pathway may consist of subtypes that vary considerably in diameter (Rockland 1995), thus affecting the conduction times. Finally, in any given cortical area, interlaminar conduction and signal processing will delay further the appearance of spikes in the neurons furthest removed from the arborizations of afferent axons. In both cat (Best et al.1986) and monkey (Maunsell and Gibson 1992) primary visual cortex, the average latency is highest in the deep and superficial layers lying most distant from the input in layer IV. Although it is probable that MT/V5 follows this same general pattern, few
studies have examined latency in MT/V5, and none have attempted to identify the specific factors giving rise to those latencies.

The extent to which signal processing and feature extraction produce delays in spike activity has not been established, and there are in fact two distinct issues here: the first is the degree to which the latency of a single neuron varies when one or more stimulus parameters are varied, the second involves the relationship between stimulus selectivities and latencies over a population of cells stimulated with their optimum stimuli. There is little question that the latency of any given neuron is affected by stimulus parameters such as orientation (Celebrini et al. 1993; Gawne et al. 1996), contrast (Gawne et al. 1996; Maunsell and Gibson 1992), size (Boltz et al. 1982), speed (Lagae et al. 1994; Lisberger and Movshon 1999), and luminance (Boltz et al. 1982; Maunsell et al. 1999). Stimulus specificity, by definition, affects response rates, and much of the effect of stimulus parameters on individual latencies may simply be due to the lower responses elicited by nonoptimal stimuli (Boltz et al. 1982; Maunsell and Gibson 1992). In this case, stimulus-latency dependencies reflect information flow through the circuitry underlying the selectivity under investigation and cannot address the larger question of the number of synapses, i.e., the circuitry itself, that may be involved in generating that selectivity, or the delay that such processing entails. At the population level, it appears obvious that the creation and elaboration of stimulus selectivities should require increasingly complex circuitry, yet a higher degree of stimulus selectivity is not invariably reflected in longer average latencies. On one hand, Nowak et al. (1995) have found that color- and orientation-selective cells in V2 have significantly longer latencies than nonselective cells, yet both those investigators and Celebrini et al. (1993) have reported that V1 cells with longer latencies have no more tendency to be orientation selective than those with shorter latencies. The degree to which latency is associated with stimulus selectivity and tuning therefore may vary according to both the type of selectivity and the cortical area where the processing takes place.

An association between latency and processing is suggested by the structure of the cortex itself. The tendency for longer latencies in cortical layers at greater removes from the input layers appears to reflect an increasing complexity in the circuitry, and cells in layers most distant from IV are indeed more likely to receive polysynaptic input (see Gilbert 1983 for review). The presence of these additional synapses will necessarily produce longer signal delays, as will conduction time through the circuitry underlying the selectivity under investigation. Once the sources of latencies are understood, then the delay between stimulus onset and the appearance of the response becomes a clue to the nature of the neural processing that may be involved in generating that selectivity, or the delay that such processing entails.

The intent of this study, then, was to investigate the extent to which the latency of MT/V5 neurons is associated with the evolution of specific receptive-field properties by examining the relationships between latency and various neuronal attributes, including stimulus selectivity, laminar distribution, and response rates. Once the sources of latencies are understood, then the delay between stimulus onset and the appearance of the response becomes a clue to the nature of the neural machinery involved in the visual process.

**METHODS**

The basic animal preparation, experimental, and testing procedures employed in this study are described in greater detail in previous reports analyzing other aspects of the present test results (Raiguel et al. 1995, Xiao et al. 1997, 1998). Single-unit extracellular recordings were made in area MT/V5 of 22 anesthetized (sufentanil; Sufenta Forte, 5 μg · kg⁻¹ · h⁻¹) and paralyzed (pancuronium bromide; Pavulon, 0.4 mg · kg⁻¹ · h⁻¹) male macaque monkeys (Macaca fascicularis) weighing between 3.2 and 5.4 kg.
Visual stimuli were circular patches of moving random dots consisting of white (48 cd/m²) dots on a dark (0.2 cd/m²) background moving coherently in the frontoparallel plane. Dots measured 0.35° in diameter with a density of 2.5 dots per square degree at the testing distance of 0.57 m. All stimuli were preconfigured and stored as sequences of 512 × 512 images on a Microvax II Workstation. Image sequences were displayed at 100 Hz using a Gould IP 9545 image computer and presented in pseudorandom order. Random dots filled the entire 25.6 × 25.6° area of the monitor at all times, but only the dots within the stimulus itself moved during presentations. Because the random dots already were present over the receptive field when motion began, motion onset coincided with the appearance of the first frame of the motion sequence.

Penetrations were made in the parasagittal plane between the superior temporal and lunate sulci, 13–17 mm lateral to the midline and at an angle of 25–30° from the vertical, pointing slightly rostrally and parallel to the superior temporal sulcus. Magnetic resonance imaging (MRI) images of individual brains generally were used to facilitate planning the penetrations. Electrolytic lesions made during the course of each penetration aided reconstruction of the electrode path and in the identification of the cortical area and layer of the recorded neurons in Myelin- and Nissl-stained sections. MT/V5 was identified on these sections by the extent of the heavily myelinated region (Ungerleider and Desimone 1986; Van Essen et al. 1981) and was readily identifiable during the experiment by the high proportion of directionally selective cells and the retinotopic organization of the receptive fields (RFs). Cortical layers were defined according to Garey (1979), but with layer III arbitrarily subdivided into three sublaminae, IIIa, b, and c, of equal thickness (Raiguel et al. 1995). Cells were stimulated monocularly, using the eye giving the stronger response. Spikes were recorded over a total of 1050 ms per presentation, including 250 ms before the onset of stimulus motion, the 300 ms of stimulus movement, and 500 ms after the stimulus had stopped. On-line analysis of responses provided feedback during the experiment.

Two quantitative tests were employed in this investigation. First the influence of the direction and speed of stimulus motion was examined using the direction test, which consisted of 48 stimulus conditions comprising 16 directions from 0 to 357.5° and three speeds of 5, 20, and 40°/s. The size of the stimulus used in this test was selected on the basis of the handplot. After the optimum speed and direction had been determined with this direction test, a two-dimensional position test (Lagae et al. 1994; Raiguel et al. 1995) was used to precisely center the stimulus display over the center of the RF before proceeding to the summation test that followed.

The second quantitative test, the summation test, examined the relationship between stimulus size and response and determined the presence and strength of any antagonistic surround. This test presented 11 concentric, circular stimuli centered on the RF and presented at the optimum speed and direction of motion for the neuron. These stimuli encompassed a range from 1.6 up to 25.6° in diameter, sufficient to cover the entire center and surround. A decrease in response as the stimulus size increased beyond a given, optimum diameter indicated the presence of an antagonistic surround. The amount of this decrease at the largest stimulus size, expressed as a percent of the maximum response, was used as a measure of the strength of that surround.

A subset (n = 66) of the cells was also tested using static gratings and edges. Contrast and luminance of these stimuli were identical to those in the motion tests and consisted of luminance edges and square-wave gratings with frequencies of 1, 0.5, 0.25, and 0.16 visual degrees (at 57 cm) presented at all orientations, encompassing a full 180° at 22.5° intervals. The stimuli were flashed onto a uniform screen of equal mean luminance following the same presentation pattern as that used for motion stimuli; a 300-ms presentation time with 750 ms between presentations. The stimulus giving the strongest response was selected for comparison with the motion response.

For statistical comparisons of responses, the response evoked by a given stimulus condition was defined as the average discharge rate during all presentations over a time period equal to the stimulus in duration but beginning at 50 ms after the stimulus onset. Spike data were analyzed as cumulative peristimulus time histograms (PSTHs) with 10-ms bins. Preliminary data analysis using a series of binwidths from 5 to 25 ms showed that the choice of binwidth had no effect on the resulting latency measurements, confirming what others have found (Nowak et al. 1995). Latencies were determined using cumulative sum analysis (Ellaway 1978), applying statistical criteria similar to those of Maunsell and Gibson (1992) and Vogels and Orban (1994) to identify response onset. First the mean and standard deviation of the spontaneous spike rate was determined from the 150-ms periods preceding stimulus onset in all runs, then the onset of the response was defined as the first bin after motion or flashed stimulus onset where the bin exceeded the spontaneous discharge rate by two standard deviations and which was followed by at least two successively increasing bins. To examine responses across the entire cell population for a given test, a population PSTH was created by combining the histograms of the individual neurons. To do so, each histogram first was normalized by setting the highest bin of the optimum condition of a test equal to 1, thus equalizing the contributions of cells with high and low firing rates.

The optimum speed of a neuron was simply the speed giving the strongest response. The preferred direction at a given speed was defined as the vector sum of the responses in all directions tested rounded to the nearest of the 16 directions. The sharpness of the tuning was expressed as the selectivity index (SI), as defined by Vogels and Orban (1994)

\[ SI = \sqrt{\frac{\sum_{i=1}^{n} S_i \cdot \sin (\alpha_i) + \sum_{i=1}^{n} S_i \cdot \cos (\alpha_i)^2}{\sum_{i=1}^{n} S_i}} \]

Where \( n \) is the number of directions tested, \( S \) is the response elicited by stimulus \( i \), and \( \alpha \) is the angle specifying the direction of motion of a given stimulus. Direction selectivity along the optimum axis was expressed as the direction index or DI (Orban et al. 1981).

RESULTS

Our sample consisted of 310 MT/V5 neurons successfully tested with the direction test, ranging in eccentricity from 0.8 to 38° with a mean of 12.6° and giving a mean response of 43 spikes/s at optimum speed and direction. All cells included in this study had an optimum response of \( \geq 10 \) spikes/s. Most (222) of these neurons also were tested with the summation test. The data in the present sample largely overlap with the 237 summation tests investigated in Raiguel et al. (1995), but data from the earliest tests (71 neurons), which used a different data format and testing sequence were not included here, whereas data from three subsequently recorded animals were added (56 neurons).

Relationship of speed and direction tuning to response latency across the population

Most of the cells tested were tuned for a given direction of motion, with a mean SI at optimum speed of 0.49 ± 0.26 (mean ± SD). Of the 310 cells tested, 65 gave their strongest response at 5°/s, 109 at 20°/s, and 136 at 40°/s. The latencies of responses to the optimum speed and direction of each neuron ranged from 35 to 325 ms, with a mean of 87 ± 45. The distribution of these latencies is shown in Fig. 1. Although the distribution is nearly symmetrical, the range is narrower than
would be expected of a true normal distribution ($P < 0.01$, Lilliefors test for normality).

There was no great variation in the response latencies among cells of different speed tunings, although the mean latency, at optimum stimulus, for neurons tuned to the slowest speed, $90 \pm 40$ (SD) ms was slightly longer than those tuned for medium ($86 \pm 45$ ms) and fast speeds ($86 \pm 42$ ms). Because response latency can depend on response strength, however (Boltz et al. 1982; Celebrini et al. 1993), it is important to assess the contribution that varying response levels may have to any observed differences in latencies. It is possible to distinguish between more meaningful differences among classes and those due to systematic variations in firing rates by examining plots of latency as a function of response rate. Any differences among classes in the curves describing that relationship indicates a specific effect of the speed preference on latency.

Figure 2A plots the log latency as a function of log response strength for all cells at the optimum speed and direction. Linear regression lines fitted to the log-transformed data for the three speed tuning classes illustrate that the relationship remains constant across these groups (slopes: $F_{2, 500} = 0.016, P \approx 1$; intercepts, $F_{2, 500} = 0.18, P > 0.5$, ANCOVA), suggesting that the longer latencies observed in slow-tuned neurons are simply due to the generally lower response levels shown by the cells of this group.

As one might expect, the preferred angle of the direction tuning bore no relationship to the latency of a neuron ($P > 0.6$, ANOVA) nor was it related in any way to the firing rate ($P > 0.8$, ANOVA). However, there was a very significant ($P < 10^{-6}$, ANOVA) inverse relationship between the width of direction tuning, as quantified by the SI and latency (Fig. 2B). Although this is to some extent explainable by a tendency ($P < 0.002$, ANOVA) for higher response rates in more sharply tuned neurons, the relationship between SI and latency remains strongly significant ($P < 4 \times 10^{-5}$, ANCOVA) even if the response is considered as a cofactor. The implication is that if the most sharply tuned neurons are those that respond most quickly, then the population tuning should broaden somewhat over time as less sharply tuned neurons begin to contribute. Figure 3A shows that this is indeed the case and that within the period between the first appearance of spike activity in the population response, at 40 ms, and the point where the maximum response rate is reached, ~80 ms, the average tuning curve becomes visibly wider. Although at least some of the broadening may simply be due to the weak nature of the initial portion of the responses, like the tip of an iceberg, it is obvious that the population response is sharply tuned for direction from its very onset. Figure 3B illustrates, for the MT/V5 population, the relationship between the evolution of direction tuning, quantified by the SI, and spike activity, here scaled so that their maxima are comparable. The SI depends on response rate and therefore rises over time: However, it can be seen that the rapid rise in SI precedes the rise in spike activity by some 10–20 ms, indicating that, initially, the SI is determined primarily by the narrow width of the tuning, but that the strength of the response gradually becomes the dominant factor. Thus the very first few spikes to appear are, in this sense, the most narrowly direction tuned.

Because direction selectivity in the two directions of the preferred axis of motion is a subset of direction tuning, it is...
unsurprising that direction selectivity expressed as the DI also
was related to response latency ($P < 10^{-6}$, ANOVA). The
trend (not shown) followed that of the SI, with higher latencies
associated with lower DIs, but because most MT/V5 cells were
generally very directionally selective, the data were much less
evenly distributed, with the majority of the DIs falling into the
80–100 range.

A consistent ($P = 0.01$, ANOVA) relationship was found
between eccentricity and latency. Neurons the receptive fields
of which were near the fovea had latencies averaging almost 20
ms longer than those located more peripherally. Closer inspection,
however, reveals that at least part of this effect is explainable
in terms of response levels, and when this covariate is
taken into consideration, the relationship is no longer significant
($P = 0.08$). Because peripheral receptive fields tend to be
tuned for higher speeds (Lagae et al. 1993), they simply reflect
the overall tendency, discussed in the following text, for higher
response rates in "fast" cells. Previous work using moving light
and dark bars also has reported higher average response rates
in peripheral neurons (Lagae et al. 1993).

**Effect of relative stimulus speed and direction on response latency within individual neurons**

Speeds slower or faster than a given neuron’s optimum simply produced longer average latencies (+4 and +5 ms, respectively), commensurate with the weaker responses, as did motion that was nonoptimum in direction (e.g., +3 and +6 ms for deviations of 22.5 and 45° from the preferred axes of motion, respectively). Others (Lagae et al. 1994; Lisberger and Movshon 1999) who have tested over wider ranges of stimulus speed, 2–50 and 0.5–100°/s, respectively, have found differences of ≤30 and 100 ms in manually measured latencies at the two extremes. However, Kawano et al. (1994) found that speed had a much more modest effect (<10 ms) on latencies of individual neurons in area MST despite the fact that MST receives direct input from MT/V5. Although the effect of the factor stimulus speed was strongly significant ($P < 0.007$, ANOVA), slopes of regression lines describing response versus latency for the two nonoptimum speeds were statistically indistinguishable from the optimum ($P > 0.5$, ANCOVA) and indicate no variation in the latency with stimulus speed that cannot be accounted for by differences in the response strength. Direction of motion produced an even stronger effect on latency ($P < 10^{-6}$, ANOVA; Fig. 4); but once again, this is an obvious consequence of stimulus tuning, and if the contribution of response strength is removed as a cofactor, the main effect, relative stimulus direction, is no longer significant ($P = 0.06$, ANCOVA).

**Latencies within the sublaminae of area MT/V5**

We also examined the response latencies in individual laminae of MT/V5 cortex. Because cells in layers most distant from IV receive largely polysynaptic input (Gilbert 1983), it is logical to assume that such polysynaptic pathways would be associated with both more sophisticated receptive-field properties and longer signal delays. Of our original sample of 310 neurons tested with the direction test, we had lamination data for 279. Of these neurons, 12 were found in layer II, 24 in IIIa, 36 in IIIb, 71 in IIIc, 75 in IV, 46 in V, and 15 in VI. The difficulty of finding and holding cells in the most superficial layers resulted in relatively low numbers of cells being recorded in layer II, and deeper layers were not always reached,
hence the central laminae tend to be somewhat overrepresented in this sample.

The latencies in MT/V5 did indeed show a distinctive laminar pattern. As one might predict on the basis of synaptic connection patterns, the overall tendency was for higher average latencies at increasing displacements from the input region around layer IV, such that lamina IIIa lags IV by 40 ms. However, layer II constituted an exception to this trend, showing a remarkably short average latency (Fig. 5A) that was statistically distinct from the adjacent layer, IIIa, at the 0.02 level (ANOVA). The uppermost lamina in fact proved distinctive with regard to a number of properties, although any conclusions must be tempered somewhat in consideration of the small size of the sample.

The next logical question concerns the origin of latency differences across layers. Is it a product of lower response rates or is there an intrinsic delay imposed by additional processing and conduction times at greater removes from the input? Response level can explain some of the effect of the variable layer because statistical significance falls from $P < 10^{-6}$ to $P < 0.002$ (ANCOVA) when response level is taken as a cofactor; but the effect is still quite significant. If the average response is plotted per layer, we see that the pattern is virtually the inverse of that shown by the latency and that responses tend to decrease in strength with increasing displacement from the input layers. Again, the uppermost lamina constitutes the exception to the overall trend with a firing rate, no higher than that in adjacent layer IIIa, that fails to mirror the much lower latency, implying that factors other than spike rates are responsible for the anomalous latency. The population PSTHs of cells tested in layers II, IIIa, and IV ($n = 9, 24, 51$; earliest data were not recorded in a format accessible to histogram analysis) are compared in Fig. 5B. Because the histograms are normalized, the influence of response rate on latency is largely obscured so that onsets in layers IIIa and IV become indistinguishable, yet a delay on the order of 10–20 ms persists between the population responses of lamina IIIa and II.

If much of the activity in layer II does indeed arise from direct subcortical input, then one consequence should be a reduced directional selectivity in both the sense of a broader directional tuning width and in the sense of directional selectivity along the preferred axis of motion because both properties are weak to nonexistent in the pulvinar and colliculus (Bender 1983; Goldberg and Wurtz 1972, Schiller 1972). We found that direction tuning, as quantified by the SI, is indeed lowest in layer II ($P < 0.02$, layer II vs. all others, ANOVA; Fig. 6A). This pattern is echoed to a certain extent by the optimum-axis direction selectivity, but laminar trends are rather less consistent (Fig. 6A). Earlier experiments that measured DIs using a single small stimulus placed in the most...
responsive part of the receptive field gave values that were higher and varied less from layer to layer but nonetheless showed a slight dip in average DI in layer II (Raiguel et al. 1995). A plot of the SI evolution over time, compiled from the averaged responses of layer II neurons (Fig. 6B), confirms that the SI reaches a maximum level only about half that of all layers combined (see Fig. 3B), but that it begins to rise at least as early as that of the remaining layers, reaching a comparatively low peak at ~70 ms. A comparison with the curve for MT/V5 as a whole (Fig. 3B) shows that this peak occurs >30 ms earlier in layer II. The population response histogram (Fig. 6B) follows a similarly early onset, with a transient component that rises to a peak at 70 ms, then quickly falls to about half its maximum value by 140 ms. This response is consistent with the sort of transient spike activity in MT/V5 that remains after a V1 lesion and apparently arises from collicular input (Rodman et al. 1989, 1990).

In view of reports that the signals that arrive at MT/V5 via pathways bypassing V1 are generated preferentially by faster motion (Hythe et al. 1995), we compared the speed tunings of layer II cells with those of other layers. We found no evidence that the short-latency neurons in layer II of MT/V5 respond preferentially to faster stimuli. The proportions of cells preferring fast, medium, or slow speeds (25, 33, and 42%) are about equal to those from the remainder of the sample (21, 34, and 45%).

Antagonistic surround and response latency

One of the well-known properties of area MT/V5 neurons is the presence of antagonistic surrounds associated with most of their receptive fields (Allman et al. 1985; Raiguel et al. 1995; Tanaka et al. 1986). It seems likely that a neuron possessing an antagonistic surround also would display a longer latency because we have speculated that surround cells represent a later stage in motion processing than nonsurround cells (Raiguel et al. 1995); thus entailing a greater number of synapses between the retinal input and the cell in question. To investigate the possible relationship between surround and latency, we first compared the latencies, as measured at optimal size in the summation test, within the two extremes of the sample: neurons in which surround antagonism produced no more than 15% inhibition at the largest stimulus (n = 22) and those in which the response was completely inhibited (n = 48) by the largest stimulus. It should be emphasized that data presented here were obtained at optimum stimulus size and that the relationship among stimulus size, surround inhibition, and latency is an additional topic that will not be taken up in the present report. The distributions of the latencies, shown in Fig. 7, clearly are shifted (P = 0.01, ANOVA) with respect to one another, with means of 66 ± 24 and 87 ± 26 ms for the low- and high-inhibition cells, respectively. Once again, however, differences in latency appear to reflect overall response levels in the two groups because the strong-surround group has a mean response rate at optimum of 32 spikes/s, whereas the group with little or no surround antagonism has a median response rate of 50 spikes/s.

Can attenuation of the response by the surround antagonism completely account for the observed differences in response latency between neurons with different levels of surround antagonism, however? To address this question, we must again examine the relationship between response and latency, this time at different levels of surround antagonism. If factors other than response rates come into play, then this relationship may be expected to differ depending on the level of surround influence. For this purpose, the sample was divided into four categories from 0 to 100% inhibition in 25% increments. Scatterplots were prepared of the log response versus log latency, and linear regressions were calculated on the log-transformed data for each of the four categories. These regression lines are depicted in Fig. 8. Statistical analysis of these regressions (ANOVA) showed that although the slopes of the relationships were not statistically distinguishable (P > 0.5), the intercepts were significantly different (0.001 < P < 0.002) and that neurons with higher levels of surround antagonism tended to have inherently longer latencies that cannot be completely attributed to the lower responses in those cells. The difference corresponds to an average increase of ~15 ms in the latencies of neurons with the strongest antagonistic surrounds (75–100% suppression) over the next-highest class (50–75% suppression).  

Laminar effects of surround on latency

As we and others (Born and Tootell 1992, Lagae et al. 1989; Raiguel et al. 1995; Tanaka et al. 1986) have reported previously, there is a marked variation in the average level of surround antagonism from layer to layer. Figure 9 summarizes the relationships among latency, response, and surround antagonism. The latency and response-level patterns across the cortical thickness reiterate those of Fig. 5A, substantiating the virtually identical results obtained in the direction tests. This figure also emphasizes the relationship that exists between the
average latency in a given layer and the corresponding surround inhibition, which follow almost identical patterns: low around the input layers and higher in the infragranular and supragranular layers, with the exception of II, where latency and inhibition are again rather low. In this uppermost layer, however, the expected concomitant rise in response rates does not occur.

Statistical analysis shows that the effect of the laminar position is indeed a significant factor ($P < 0.02$, ANOVA) with respect to latency. If the surround inhibition is considered as a cofactor, however, then the effect of laminar position is no longer significant ($P > 0.20$) nor is it significant if response is considered as a cofactor ($P > 0.30$). This suggests that laminar effects are largely a consequence of response levels, which in turn are the result of varying levels of surround antagonism in the different layers. This idea receives some support from the finding that the inhibition class (Fig. 8) significantly affects response levels ($P < 0.01$, ANOVA) and implying that the strength of the inhibitory surround somehow remains a factor in determining the response level, despite the use of optimal-sized stimuli in the testing procedure.

Comparison of responses with static and moving stimuli

Because our comparisons of latencies assume a consistent relationship between response strength and latency, it is logical to wonder how general this relationship might be and whether the same relationship might hold for a completely different sort of stimulus, e.g., a flashed, static grating or edge. Because many experiments often are performed on the same units, 66 of the earliest neurons in our data set also had been tested using static stimuli, and 56 of these gave measurable responses to one or more of the static stimuli. Responses to the optimum static stimuli produced average latencies some 5 ms shorter ($P < 3 \times 10^{-6}$, paired $t$-test) than those to moving stimuli. No discernable relationship was found between the latencies as determined with the two types of stimuli ($R^2 < 10^{-2}$). Although our initial assumption had been that the relationship between response and latency was a universal, Poisson phenomenon, we were surprised to learn that the latency for the flashed stimulus is more or less constant, with a log-log slope of only $-0.014$ (Fig. 10) compared with the corresponding slope for the motion stimulus of $-0.26$ ($F_{2,100} = 4.8, 0.01 < P < 0.02$, ANCOVA).

**DISCUSSION**

**Significance of latencies and sources of variation**

The latency of neuronal responses measured in the MT/V5 population varied with the evoked discharge rate in a relation-
ship that remained consistent over a wide range of response rates for a given stimulus type (Figs. 2 and 8), as has been observed in retinal ganglion cells (Boltz et al. 1982). It has been emphasized that there is no single value, however determined, that can adequately represent the absolute latency of a neuron (Nowak and Bullier 1997). This makes comparisons between studies difficult, yet comparisons of latencies across stimulus parameters within a study nonetheless can be meaningful for a given set of criteria. Indeed, it should be pointed out that latency-response relationships in general may be highly stimulus-specific, as our comparison of responses to flashed and moving stimuli (see last section of Discussion) would indicate.

The average latency of 87 ms in the present study is in good agreement with the 94 ms previously reported in MT/V5 using moving bars (Raiguel et al. 1989), but longer than the 72 ms recently found (Schmolesky et al. 1998) using flashed bars which typically produce transient responses with shorter rise times and shorter latencies (Maunsell 1987; Nowak and Bullier 1997). Although ranges (10–90th percentile) of $\approx 100$ ms are common for extrastriate areas (see Nowak and Bullier 1997 for review), a narrower range might be expected for area MT/V5, considering its restricted input (Maunsell et al. 1990, Movshon and Newsome 1996; Shipp and Zeki 1989a,b). The range of latencies in our experiments, only 80 ms if expressed as the 10–90th percentile range, suggests area MT/V5 indeed lies toward the lower end of the spectrum for extrastriate cortex.

The existence of a homogeneous input eliminates at least one source of variability (Nowak and Bullier 1997), making MT/V5 ideal for investigating the remaining variables that are associated with the polysynaptic nature of the signal processing itself, such as synaptic delays, integration time, and feedback from other cortical areas.

One obvious source of variation in the observed latency is certainly the number of routes by which input may reach MT/V5. Although the number of cells recorded in the most superficial layers precludes any definitive conclusions, all the properties of these neurons that were investigated, including short latencies (Finlay et al. 1976), laminar position (Benevento and Rezek 1976), and direction selectivity (Bender 1983; Goldberg and Wurtz 1972; Schiller 1972) are consistent with a collicular input. Moreover, these distinctions achieve statistical significance despite the small numbers of recorded cells. Lesion studies (Rodman et al. 1989, 1990) have confirmed that area MT/V5 receives a fairly substantial input from the superior colliculus, implying a high probability of encountering neurons receiving such input. Collicular neurons are also poorly tuned for the axis of motion, and we found that directional tuning for a particular axis of motion is correspondingly weakest in layer II. However, the responses that remain after striate lesions (Rodman et al. 1989) or reversible inactivation (Girard et al. 1992) retain much of their direction tuning, leading to speculation that this selectivity might be generated, or at least refined, within MT/V5 itself (Girard et al., 1992; Gross 1991; Rodman et al. 1989, 1990). Although the early onset of direction tuning in layer II neurons (Fig. 6B) suggests that there is some degree of direction selectivity already present in the input, this is nonetheless much weaker than in deeper layers, the neurons of which may well act to sharpen that tuning.

There are other sources of rapid-onset neurons in MT/V5, and the vast majority of early spike activity outside layer II probably arrives via more conventional intercortical pathways. One such source of short-latency input may be the direct afferents from V1 having large fibers and boutons (Rockland 1989) and axonal conduction times on the order of $\approx 2$ ms (Movshon and Newsome 1996). This input appears to be confined to layers 3, 4, and 6 (Rockland 1989) and thus could not account for the short-latency spike activity observed in the layer II, but the extensive arbors of these axons may provide the basis for early spike activity in the deep and middle layers.

Visually evoked potentials measured in human subjects have suggested that fast-moving stimuli ($\approx 22$’s) activate MT/V5 first (ffytche et al. 1995), whereas slow-moving stimuli ($\approx 6$’s) initially activate V1. Those investigators concluded that slow stimuli are processed by a pathway that includes V1, whereas faster stimuli use a separate pathway bypassing V1, implying that cells with faster tunings in MT/V5 should have shorter latencies. Our failure to find faster speed tunings associated with layer II neurons suggests that the short-latency input into this layer that we observe cannot be this proposed pathway. Although we did find an association between shorter latencies and faster speed tunings in MT/V5 as a whole, the shorter latencies appeared to be explainable on the basis of response strengths. Because the majority of cells in MT/V5 respond best to faster speeds (Kawano et al. 1994; Lagae et al. 1993), faster motion will produce higher response rates with correspondingly shorter latencies. V1, with a high proportion of cells tuned for speeds $\approx 10^3$’s (Orban et al. 1986), would respond poorly at faster speeds while responding to slow stimuli vigorously and with correspondingly short latencies. Moreover, the layers in V1 that project to MT/V5 have been found to contain few low-pass cells (Orban et al. 1986), so that slow stimuli presumably would elicit only modest responses, with longer latencies, in area MT/V5. Thus the results reported by ffytche et al. (1995) are also explainable on the basis of the speed-response curves of V1 and MT/V5 without the necessity of evoking separate pathways.

**Antagonistic surround and latency**

The existence of a relationship between the level of surround inhibition expressed by a neuron and the latency of its response might not be unexpected because the former, by definition, sharply affects response. However, the data presented here were measured using stimuli of optimum diameter that presumably include little or none of the surround. Moreover, even if there was overlap between center and surround regions, such that the surround exerted an influence on the response levels, and hence, latency, that influence should be accounted for and factored out by analysis of covariance, which was not the case. The 15-ms increase in response latency associated with the presence of a surround thus appears to be an intrinsic property of surround neurons in MT/V5 and suggests that neurons with strong, well-developed surrounds may represent a later stage in processing than those which have no or only weak surrounds. Such neurons presumably could be created either by combining intracortical input from lower-order neurons or from other nearby neurons at the same hierarchical level or by combining local representations of receptive fields with feedback from higher areas (Tanaka et al. 1986).

Significantly longer onset latencies in the surround com-
pared with the center would favor feedback as the source of the antagonistic surround. Previous studies examining antagonistic motion surrounds and latency in MT of the owl monkey have reported that the onset of the surround antagonism began <40 ms (the bin size in the experiment) later than that of the center response (Allman et al. 1985). Later studies suggest that the difference is probably no more than 10–15 ms in the macaque (Orban 1998; Raiguel et al. 1998), corresponding to one or two intervening synapses and suggesting that the surrounds are created by combining signals from within MT/V5 itself rather than being imposed by feedback from higher areas.

Neurons in area V1 of the macaque also possess antagonistic surrounds. These react to stimulus qualities such as orientation, texture, color, luminance, and disparity (Knierim and Van Essen 1992; Sillito 1995) and tend to suppress responses when the stimulus in the RF matches that of the surround, in a manner analogous to the way antagonistic surrounds react to motion in area MT/V5. Onset delays ranging from 7 to 50 ms with respect to response onset have been reported for surround influences in V1 (Knierim and Van Essen 1992; Lee et al. 1998; Zipser et al. 1996), but no correlation has been reported in V1 between the strength or presence of such surrounds and the latency of the response (Knierim and Van Essen 1992) as we have found in MT/V5. Although some investigators have reported that many of the modulatory effects elicited by the surround can disappear under anesthesia (Lamme et al. 1998), suggesting feedback from higher areas, others (Hupe et al. 1998) have shown, through inactivation studies, that such feedback to V1 from MT/V5 largely amplifies responses in V1, rather than inhibiting responses as antagonistic surrounds do and arguing that feedback, from MT/V5 at least, does not give rise to V1 surrounds. Perhaps surrounds in V1, like the neurons themselves, represent a more heterogeneous population than in MT/V5, with some generated by feedback from V2 or higher areas, whereas others arise locally through lateral or feedforward connections.

**Laminar influences**

With the exception of layer II, as discussed in the preceding text, the distribution of latencies across layers closely follows that described in V1 by Maunsell and Gibson (1992): lowest in the input layers and slowly rising with increasing vertical displacement from layer IV. The generality of this distribution is demonstrated by cat primary visual cortex, which follows a similar pattern save that in that species, afferents into layer V1 reduce average latencies in this layer to levels approaching that of IV (Best et al. 1986). The increase in latencies observed across the thickness of the cortex probably has its rather straightforward origin in the polysynaptic input to the more superficial layers (Levitt et al. 1996), and the synaptology of MT/V5 almost certainly follows a similar pattern. Each neuron in the sequence will add ~5 to 10 ms of integration time (Nowak et al. 1995; Nowak and Bullier 1997), so that the 20-ms delay in activity in the upper layers (Fig. 9) would correspond to two or four intervening synapses. This is similar to what has been reported for areas V1 and V2, both in terms of latency (Maunsell and Gibson 1992; Nowak et al. 1995) and synaptology (Levitt et al. 1996).

Much of the latency increase associated with the upper laminae may be attributable differences in response levels, however, and thus it is not obvious how much may be due to synaptic delays and conduction time per se and how much may simply be due to lower response levels. Yet this need not be a simple either/or proposition but simply may represent two aspects of the same phenomenon. Lower response levels in fact could be a byproduct of passing the information from neuron to neuron, particularly if the stimulus specificities of the classical receptive field are generated largely through inhibitory mechanisms as many have suggested (Bishop et al. 1971; Bonds 1989; Ferster and Lindström 1983; Sillito et al. 1980; Wörgötter and Eysel 1991).

A second element that may provide the link among layer, response, and latency is the evolution of response properties involving additional selectivities for parameters not specifically tested here, such as depth, orientation, or disparity. Recent evidence suggests that the surround configurations may be more complex than previously suspected (Xiao et al. 1995, 1997, 1998) and that they are capable of specifying more sophisticated stimulus properties, such as the direction of a speed gradient, that our testing procedure did not consider. In other words, the generally stronger surround antagonism in neurons at more advanced stages of processing may parallel an increase in the selectivity of those neurons for specific, but unknown stimulus characteristics with a consequent decline in response levels. In this regard, any “standard” stimulus will produce a range of response levels, and hence latencies, depending on the degree to which it matches these unknown specificities. A second consequence of these emergent selectivities will be an increased overall scatter in the latencies of any given layer because the stimulus may or may not match the tuning of a particular neuron for those properties, as chance dictates.

**Computational issues and direction tuning**

The strong direction tuning from the very onset of the spike trains indicates that MT/V5 neurons should have the capacity to specify the direction of motion in even the earliest part of the response. It has been found using information theory (McClurkin and Optican 1996; Toveé et al. 1993) that the information available during the first 20–50 ms of firing is sufficient to specify most of the information carried by the spike train. The availability of such information is reflected in the rapid rise in the SI, which actually precedes the rise in spike rate observed in our sample. The extension of the spike period beyond this initial discharge increases the overall information content of response (Toveé et al. 1993), as shown by the fact that the SI continues to rise despite the slight broadening in the directional tuning width. The initial sharply tuned but statistically weak portion of the signal corresponds to the “fast brain” aspect of the neural circuitry (Nowak and Bullier 1997), comprising those processes that depend on precise temporal relationships and require rapid conduction and processing, whereas the later part of the response, where distinctions between responses to optimal and nonoptimal stimuli are maximal (Oram and Perrett 1992), differentiate complex spatial or spatiotemporal patterns using feedback circuits and entailing longer latencies (Maunsell 1987). Temporal “smearing” of the response, moreover, permits interaction with other neurons higher up in the processing hierarchy and provides an opportunity for additional...
stimulus specificities to evolve (Knierim and Van Essen 1992) and for finer discriminations to take place (Zohary et al. 1990).

Like orientation tuning (Celebrini et al. 1993; Ringach et al. 1997; Somers et al. 1995), direction selectivity could arise from feedforward mechanisms or could additionally involve recurrent intracortical feedback (Maex and Orban 1996; Murphy and Humphrey 1999). Feedforward models emphasize convergence or synchronization of input (Gawne et al. 1996; Maunsell and Gibson 1992; Nowak and Bullier 1997) onto neurons that behave as coincidence detectors (König et al. 1996), such that those sharing similar tunings for a given characteristic are mutually reinforcing (Löwel and Singer 1992; Toyama 1988). This model can just as readily apply to MT/V5 because V1 input is already directional (Movshon and Newsome 1996) and would account for the tendency for higher firing rates to be associated with sharper tunings. Combining slightly different optima to create a broader tuning would mean that the stimulus is not optimal for some components, resulting in a signal that is not only weaker from the outset but is relatively desynchronized due to the different response latencies of the components. On the other hand, rapid, local intracortical feedback could further sharpen direction tuning in MT/V5 through excitatory connections from layer VI onto layer IV neurons (Grieve and Sillito 1991), producing tuning that develops over a very short time course and firing rates that are highest in input layers and in cells that are more sharply tuned. The amplification of layer IV responses need not necessarily come from neurons in other layers but even could be provided by other afferent axons (Rockland 1989, 1995) in a feedforward arrangement. Such a mechanism has the additional benefit of reamplifying the signal at each succeeding cortical area and would result in the laminar response patterns we observe.

Latencies in static versus moving stimuli

Analysis of data comparing flashing and moving stimuli in MT/V5 and preliminary work in V1 and V2 (unpublished results) indicate that the latencies of responses to these two types of stimuli differ significantly in their relationships to the strength of those responses. Moving stimuli reveal a dependence on response strength in all three areas that is largely or entirely lacking using flashed stimuli. Others have reported that in V1, average latencies to such stimuli appear to remain constant across cells (Richmond et al. 1997), at least when neurons are tested with their optimum stimulus (Celebrini et al. 1993). Part of the distinction between flashed and motion responses may lie in the mechanics of stimulus detection. A flashed stimulus can be registered by input from a single retinal locus, whereas detection of movement necessarily comes from neurons in other layers but even could be provided by other afferent axons (Rockland 1989, 1995) in a feedforward arrangement. Such a mechanism has the additional benefit of reamplifying the signal at each succeeding cortical area and would result in the laminar response patterns we observe.

Like orientation tuning (Celebrini et al. 1993; Ringach et al. 1997; Somers et al. 1995), direction selectivity could arise from feedforward mechanisms or could additionally involve recurrent intracortical feedback (Maex and Orban 1996; Murphy and Humphrey 1999). Feedforward models emphasize convergence or synchronization of input (Gawne et al. 1996; Maunsell and Gibson 1992; Nowak and Bullier 1997) onto neurons that behave as coincidence detectors (König et al. 1996), such that those sharing similar tunings for a given characteristic are mutually reinforcing (Löwel and Singer 1992; Toyama 1988). This model can just as readily apply to MT/V5 because V1 input is already directional (Movshon and Newsome 1996) and would account for the tendency for higher firing rates to be associated with sharper tunings. Combining slightly different optima to create a broader tuning would mean that the stimulus is not optimal for some components, resulting in a signal that is not only weaker from the outset but is relatively desynchronized due to the different response latencies of the components. On the other hand, rapid, local intracortical feedback could further sharpen direction tuning in MT/V5 through excitatory connections from layer VI onto layer IV neurons (Grieve and Sillito 1991), producing tuning that develops over a very short time course and firing rates that are highest in input layers and in cells that are more sharply tuned. The amplification of layer IV responses need not necessarily come from neurons in other layers but even could be provided by other afferent axons (Rockland 1989, 1995) in a feedforward arrangement. Such a mechanism has the additional benefit of reamplifying the signal at each succeeding cortical area and would result in the laminar response patterns we observe.

REFERENCES
