

Chapter 5

Attention and Signal Detection: A Practical Guide

Vincent P. Ferrera

A faint tap per se is not an interesting sound; it may well escape being discriminated from the general rumor of the world. But when it is a signal, as that of a lover on the window-pane, it will hardly go unperceived.

– William James [1]

5.1 Detection of Weak Signals

The ability to detect weak signals in the environment can have a profound impact on an organism's ability to survive and reproduce. This aspect of perception is therefore likely to have been optimized by natural selection. Part of this optimization may involve strategies to maximize performance by allocating scarce neural resources. The ability to allocate limited resources by selecting and prioritizing sensory information is often what is meant when people talk about selective attention [2]. The notion of a limited capacity filter has been invoked to explain why orienting attention to a particular location in space or a particular stimulus feature enhances detection and shortens response times. This view has given rise to imaginative metaphors such as the “spotlight” [3, 4] or “zoom lens” [5] of attention. An alternative, albeit less poetic, view considers attention from the standpoint of a decision-maker trying to make sense of noisy signals arising from multiple detectors. In this view, what is commonly referred to as “attention” may be a manifestation of the effect of uncertainty on the behavior of an ideal observer [6, 7]. While precise definitions are elusive, it is reasonable to say that attention includes a collection of computational strategies that enhance the detection and discrimination of weak signals and/or refine the behavioral response to such signals. These strategies might include increasing the signal-to-noise ratio of individual neurons, optimizing decision parameters, and identifying subsets of detectors (e.g., neurons) that are more reliable.

V.P. Ferrera (✉)

Department of Neuroscience, Columbia University, 1051 Riverside Drive, Unit 87, New York, NY 10032, USA

e-mail: vpf3@cumc.columbia.edu; vincent.ferrera@gmail.com

Stimulus detectability across the visual scene is one way to quantify perceptual salience. Salience maps are important for both human and machine vision systems as they indicate areas of heightened interest, attention, and action. Signal detection theory provides various strategies for computing salience maps. Salience maps computed using principles of signal detection can incorporate the effects of prior information (i.e., environmental cues or knowledge about target prevalence), observer bias, and/or economic value.

The conversion of sensory signals into percepts, decisions, and actions occurs over multiple stages of neural processing. At which level does attention act? Does attention affect the quality of incoming sensory information? Does it affect decision-making, response selection, or even later processes? This issue is part of the long-standing debate over “early” vs. “late” selection of signals [2, 8].

The idea behind early selection is that multiple stimuli compete for attention at an early stage of sensory processing. Attention biases this competition by enhancing the representation of behaviorally relevant stimuli [9]. Thus, when attention is directed toward a particular location or object, it improves the quality of sensory data acquired at the focus of attention. Improved quality means that the neural representation has higher fidelity, stronger signal, less noise. Better signal to noise should enhance the detectability of weak signals.

The “late” selection hypothesis holds that attention acts mainly on higher-order processes, leaving sensory representations largely intact. For example, attention can act at the level of response selection by adjusting decision criteria. An observer may have prior information that creates an expectation that the stimulus will occur at a particular time or place. This could be due to statistical regularities in the environment or to the presence of reliable cues, either natural or artificial. If a stimulus is expected to occur in a given place or at a given time, the observer may require less sensory evidence to report that it was present. Thus, they may lower their internal decision threshold or adopt a bias in favor of making a positive response. The expectation of a stimulus does not necessarily mean that the quality of the sensory evidence provided by that stimulus is better, but rather that the prior likelihood of the stimulus biases the observer to report that it is present.

To understand how attention might improve an observer’s performance, it is useful to introduce the framework of signal detection theory (SDT [10, 11]). SDT provides a simple set of computations to select responses based on factors such as signal strength, stimulus probability, and the consequences of different responses. The underlying model for SDT is that signals in the environment cause changes in the internal state of an observer. Changes in internal state then guide categorical responses, such as “yes/no” or “seen/not seen.” The SDT model affords a great deal of flexibility in mapping stimuli onto responses. Flexibility derives from the probabilistic relationship between external signals and internal states, and the criterion-dependent relationship between internal states and responses.

SDT clarifies the distinction between *stimulus detectability* and *response bias*. Detectability is a function of the sensory signal alone. It is the certainty with which an external event in the environment can be inferred from the internal state of the observer. Detectability depends only on the difference in the observer’s internal state

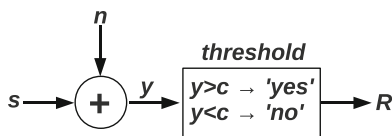


Fig. 5.1 Signal detection model. Noise, n , is added to external signal, s , to produce an internal state, y , which is compared to criterion, c , to determine response, R

Table 5.1 Confusion matrix for signal detection model

	Stimulus present	Stimulus absent
Respond “Yes”	<i>Hit</i>	<i>False alarm</i>
Respond “No”	<i>Miss</i>	<i>Correct rejection</i>

when the external stimulus is present as compared to when the stimulus is absent. It does not depend on the relationship between the observer’s internal state and their response, which may be biased toward one alternative or the other, independently of the signal.

Formally, the premise of signal detection theory is that the internal state of an observer (y) is perturbed by an external signal (s) that is affected by noise (n). In the simplest case of additive noise, $y = s + n$ (Fig. 5.1). The internal state, y , is then compared to a threshold decision criterion, c , to generate a binary response of “yes/no” or “seen/not seen.” The detectability of the stimulus is entirely determined by the characteristics of the signal and noise. However, the response can be biased depending on the level of the decision criterion. A change in criterion might cause the observer to report that the stimulus is present more or less often even though there is no real change in stimulus detectability.

In the most basic case, the stimulus takes one of two values (present or absent), and the response also has two possible values (yes, no). There are thus four possible outcomes (Table 5.1). Hits and correct rejections are both correct responses. Misses and false alarms are incorrect.

When the stimulus is present, it gives rise to an internal state drawn from a probability density function called the signal distribution (S). This distribution consists of signal + noise, as noise is always present. When the stimulus is absent, the internal state is drawn from the noise distribution (N). Internal state might correspond to the instantaneous firing rate of a neuron, or the number of action potentials fired in a specified time interval (spike count). Figure 5.2 (left panel) shows hypothetical examples of signal (blue bars) and noise (gray bars) distributions. The dashed vertical lines represent two different criterion values. For a given criterion value, the *hit rate* is the proportion of the signal distribution that is greater than the criterion. Similarly, the *false alarm rate* is the proportion of the noise distribution that is greater than the criterion. The miss rate is 1.0, hit rate, and the correct rejection rate is 1.0, false alarm rate. Observer performance is completely characterized by the rates of hits and false alarms.

To obtain a criterion-independent estimate of detectability, one can vary the criterion level through the entire range of states represented in the signal and noise

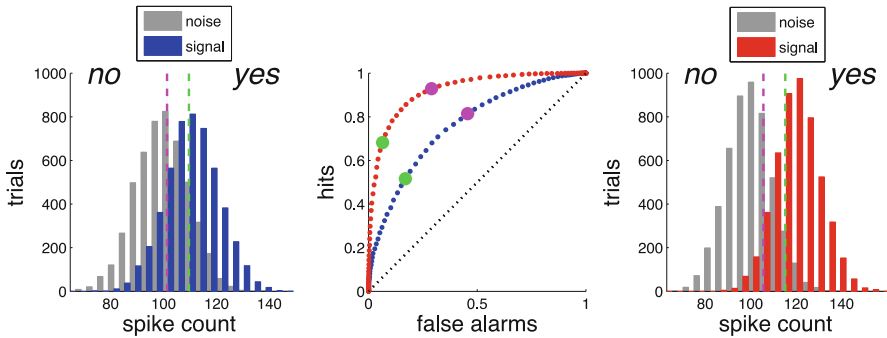


Fig. 5.2 (Left) Spike count histograms for stimulus present (blue) or absent (gray) trials. Dashed vertical lines represent different criterion values. (Middle) ROC curves. “Hits” are the number of hits divided by hits + misses. “False alarms” is the number of false alarms divided by false alarms + correct rejections. Blue dots are the ROC derived from the distributions in the left panel. Red dots are the ROC derived from the distributions in the right panel. Green and magenta dots are the points of the ROC curves corresponding to the criterion levels shown in the left and right panels. (Right) Histograms for a case with stronger signal

distributions. For each criterion level, the hit rate can be plotted against the false alarm rate (Fig. 5.2, middle). The resulting curve is called the “receiver operating characteristic” or ROC curve. The area under the ROC curve (AROC) is a measure of stimulus detectability across all criteria. Imagine drawing two random samples: one from the signal and one from the noise distribution. The area under the ROC curve is the probability that the sample drawn from the signal distribution is larger. In psychophysics, the ROC area is equal to the percentage of correct responses for an ideal observer in a two-interval forced choice experiment.

Based on how the ROC curve is constructed, it follows that changing the decision criterion, e.g., from the magenta to the green line in Fig. 5.2, only moves one along the ROC curve. The magenta and green dots superimposed on the blue curve in Fig. 5.2 (middle) are the hit and false alarm rates corresponding to the criterion levels in the left panel. Changing the criterion does not change the ROC curve itself. To do this, there must be a change in the amount of overlap between the signal and noise distributions. The right panel in Fig. 5.2 shows the distributions for a stronger signal with the same noise as in the right panel. The corresponding ROC curve is shown in red in the middle panel. The increased area under the red curve means that the signal can be more reliably detected.

It should be clear that stimulus detectability depends on the overlap of the signal and noise distributions, which in turn depends on two factors: the separation between the means of the signal and noise distributions, and the variances of those distributions. Attention can therefore improve detectability by increasing the former and/or reducing the latter. In SDT, these are the only two variables that affect the internal representation of signal quality. However, attention may also act by optimizing the decision criteria that determine the observer’s response.

5.2 Effect of Stimulus Probability

The likelihood that a stimulus will occur during a given observation period is referred to as its prior probability. In an unbiased experiment, stimulus-present and stimulus-absent trials should occur in equal proportion, so that the prior probabilities of signal and noise are both 0.5 and the percent correct that can be achieved by guessing is 50 %. In most real-world environments, the signal and noise probabilities are not necessarily equal. Rare or novel signals may attract attention by an oddball effect. Frequent signals may result in sensory adaptation, thus weakening their internal representation. Response habituation can also play a role; if a stimulus is quite rare, then observers may fall into a habit of responding “no.” This habitual response may cause observers to miss a rare stimulus if they are not vigilant. Some studies report that prevalence effects can result in miss rates of up to 50 % [12]. Other studies with medical images (chest x-rays) have reported that prevalence effects are negligible [13].

Observers can take advantage of variations in stimulus probability by adapting their decision criteria. These adjustments can be made without any explicit knowledge about stimulus probability itself. SDT naturally handles cases where stimulus probability is different from 0.5. In the unbiased case, the areas of the signal and noise probability distributions are both equal to 0.5. If there is a preponderance of signal trials, then the area of the signal distribution will be between 0.5 and 1.0, while the noise distribution will have area = 1.0 – signal area.

To understand how prior probability affects detectability, it is important to realize that the hit and false alarm rates are conditional probabilities. Specifically, hit rate is the probability that the internal state, y , is greater than the criterion, c , given that the stimulus is present: $p(y > c \mid S)$. Likewise, false alarm rate is conditioned on the absence of the stimulus, thus $p(y > c \mid N)$. Because hit rate is conditioned on the presence of the stimulus, changing the likelihood that the stimulus is present does not change the hit rate. For a given criterion level, the proportion of the signal distribution that is greater than the criterion is invariant to scaling of the distribution. The same goes for the false alarm rate. Hence, stimulus probability has no effect on the ROC curve and thus no effect on stimulus detectability.

Figure 5.3 illustrates this by showing cases where the signal has a low probability (left panel) or high probability (right panel). The ROC curves are the same for both conditions (middle). Another way to think about this is that the ability to detect a stimulus depends only on the strength and fidelity of its representation in the nervous system at the time the stimulus is present. It does not depend on the past history of the stimulus. This surprising feature of ROC curves is advantageous in areas like medical diagnosis because the probabilities of S and N are generally unknown (and difficult to measure). Thus, ROC curves are understood to provide a reliable metric for diagnostic efficacy independent of the relative prevalence of S .

It may seem counterintuitive that detectability is not affected by stimulus probability. Certainly, prior knowledge about the signal must confer some performance advantage, and it does. But the advantage derives from the fact that observers are

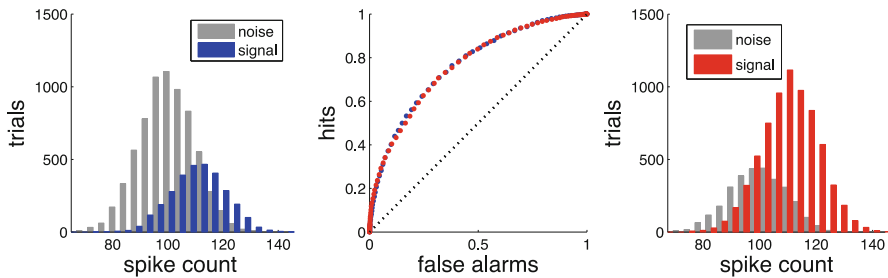


Fig. 5.3 (Left) Histogram of spike counts for stimulus-present (*signal*) and stimulus-absent (*noise*) trials with low stimulus probability. (Middle) ROC for (blue dots) and high (red dots) stimulus probability. (Right) Histograms for high stimulus probability

able to improve performance by altering their decision strategy. If the stimulus is more likely to appear than not, then there is an advantage to using a more liberal decision criterion (smaller value of c) for responding “yes.” In this case, simply closing one’s eyes and guessing that the stimulus is present or saying that it is always present will yield performance greater than 50 % correct. Signal detection theory can be used to determine the value of the decision criterion that will optimize performance (percent correct) for a given signal strength and probability. Analytically, the optimum criterion is the value of y that satisfies the following equation:

$$p(S) * f_s(y) = p(N) * f_n(y)$$

where $p(S)$ and $p(N)$ are the signal and noise prior probabilities (signal present or absent, respectively) and $f_s(y)$ and $f_n(y)$ are the unweighted signal and noise probability densities (see [14] for derivation).

Figure 5.4 illustrates simulations where the signal has low (top left) or high (top right) probability. Performance, in terms of percent correct detection, is plotted as a function of criterion level in the bottom left for the case of high (red) or low (blue) signal probability. The dashed vertical line indicates the criterion level that optimizes performance. The optimum criterion can be computed for any signal probability (bottom right). These simulations show that, while knowledge of stimulus probability does not affect detection, performance may nevertheless be improved by selecting the optimal decision criterion.

One of the chief complaints about SDT is that it seems to assume that the probability distributions are known with arbitrary precision. In practice, observers may not know the shapes of these distributions or the prior probabilities of signal and noise. However, there are simple iterative algorithms for adapting the decision criterion that produce near-optimal performance and are based only on quantities available to the observer, for example, their behavioral response and the feedback they receive (assuming feedback is given). First, note that Table 5.1 can be rearranged as follows (Table 5.2):

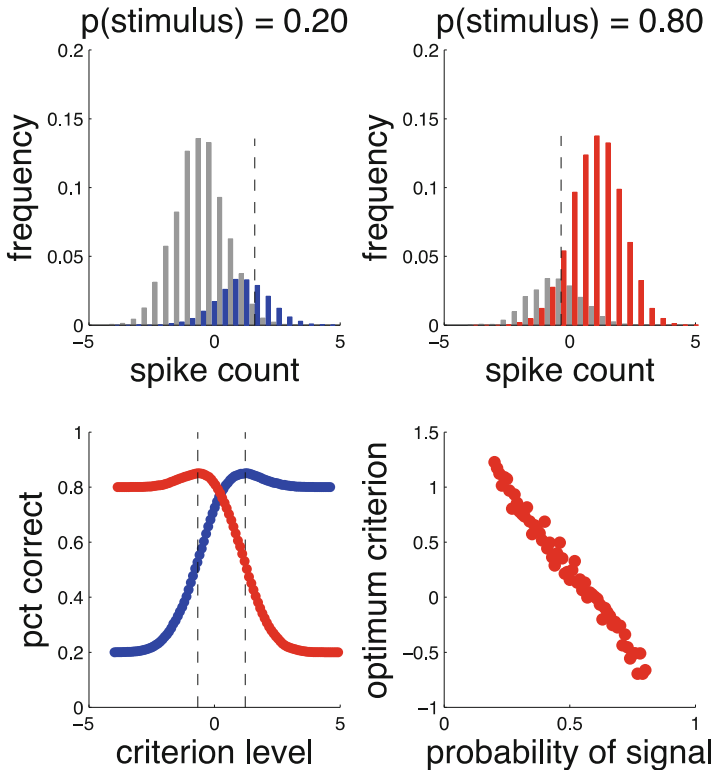


Fig. 5.4 (Top row) Spike count histogram for low (blue) and high (red) stimulus probabilities. (Bottom row) Performance as a function of criterion level and optimum criteria for different signal probabilities

Table 5.2 Confusion matrix for signal detection model reordered by response outcome

	Correct	Incorrect
Respond “Yes”	<i>Hit</i>	<i>False alarm</i>
Respond “No”	<i>Correct rejection</i>	<i>Miss</i>

Hence, an observer can deduce whether the stimulus was present (hit, miss) or absent (false alarm, correct rejection) based on the conjunction of their response (yes/no) and the outcome (correct/incorrect). The observer can use this information to estimate the prior likelihood of the stimulus. Knowing only their response and the outcome, the observer can optimize their decision criterion based on feedback.

An iterative algorithm for optimizing the decision criterion is the following: (1) after each “yes” response, the criterion level is incremented in proportion to the rate of signal-absent trials and, (2) after each “no,” the criterion is decremented in proportion to the rate of signal-present trials. This can be quantified by the following updating rules:

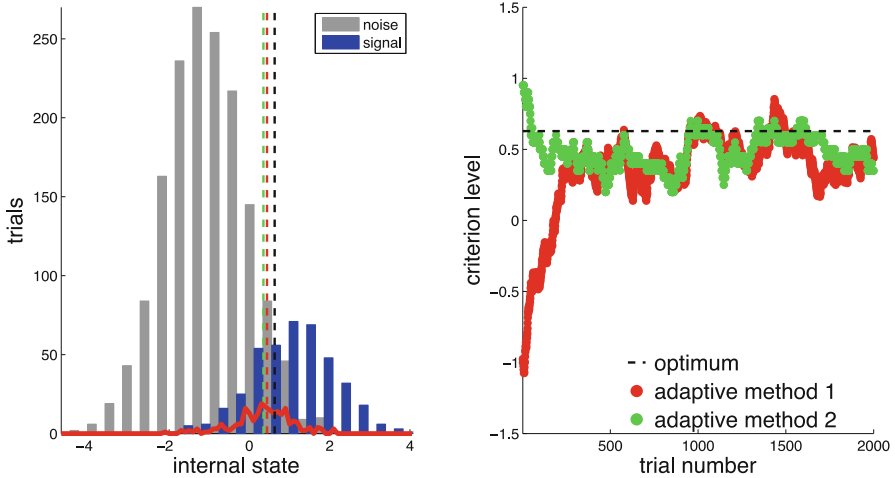


Fig. 5.5 Optimization of decision criterion. (Left) Signal and noise distributions with signal probability = 0.2. Red and green vertical dashed lines are final criterion values for two adaptive procedures. Black dashed line is optimal criterion. Red line is the distribution of error trials. (Right) Adaptive criteria as a function of trial number

1. If response = “yes,” $c(n + 1) = c(n) + k * p(N)$
2. If response = “no,” $c(n + 1) = c(n) - k * p(S)$

The criterion level on trial n is denoted by $c(n)$, k is the learning rate, $p(N)$ is the probability that the stimulus was absent, and $p(S)$ is the probability of stimulus present. These probabilities are not known in advance, but are continuously updated based on feedback. This method is stable and converges to the criterion value corresponding to an internal state that is equally likely for stimulus-present and stimulus-absent trials, i.e., $p(y | S) = p(y | N)$. An example of the algorithm’s performance is shown in Fig. 5.5 (method 1).

If feedback (“correct” or “incorrect”) is given after each trial, then the observer can deduce whether the outcome is a hit, miss, FA or CR, and thus whether or not the stimulus was present. Therefore, the information required to implement this procedure is available to the observer and does not require prior knowledge of stimulus probability.

There is another algorithm for adapting the decision criterion that uses only feedback on error trials. Specifically, after each false alarm, the criterion is incremented a small amount. After each miss, the criterion is decremented by the same quantity. This process converges on the criterion for which the miss and false alarm rates are equal, which is close to the optimum criterion for minimizing the error rate (Fig. 5.5, method 2). It is easy to show that this algorithm is stable: if the criterion value is too high, then misses outnumber false alarms and the criterion is decremented until the miss rate equals the false alarm rate; if the criterion is too low, false alarms outnumber misses resulting in a net increment. Furthermore, as long as the signal and noise distributions have positive area, there is always a criterion value

for which misses = false alarms. This can be demonstrated by considering that, as the criterion goes from $-\infty$ to ∞ , the false alarms start at a finite, positive value and then decrease to zero. At the same time, misses increase from zero to a finite positive value. Therefore, the miss and false alarm curves must cross. This procedure is therefore guaranteed to converge and has the additional advantage that it does not rely on an estimate of stimulus probability. Procedures based only on correct responses can also be used, but tend to converge at criterion values that are far from the optimum.

5.3 Effect of Costs and Benefits for Various Outcomes

SDT has four classes of outcome: hits, misses, false alarms, and correct rejections. In real-life situations, each outcome has an associated cost or benefit. Misses and false alarms are both incorrect outcomes, but are not equally costly. If a person has a medical exam, the cost of a false negative (miss) might be that they will not receive treatment and their condition may worsen. The cost of a false positive is that they could receive a treatment that is unnecessary. One of these outcomes may be catastrophic (e.g., an infection that becomes life threatening), while the other is relatively benign (taking a superfluous course of antibiotics). Likewise, the benefit of a hit may be greater or less than a correct rejection.

If one can assign a numerical value to each outcome, then there is a formula for the expected value (EV) of each trial [14]:

$$EV = V_h * p(h) + V_m * p(m) + V_{fa} * p(fa) + V_{cr} * p(cr)$$

where V is value, p is probability, and the subscripts denote the various outcomes. The values of misses and false alarms are typically negative as these outcomes represent costs. This formula also incorporates effects of prior stimulus probability as this affects the outcome probabilities.

Figure 5.6 shows an example of the effect of value-weighted outcomes on the optimal decision criterion. In the left panel are weighted outcomes for hits (green), correct rejections (blue), misses (black), and false alarms (red). The stimulus probability is 0.5. The dashed lines represent the balanced case where correct responses have a value of 1 and errors have a value of -1 . The expected payoff and optimum criterion is shown in black in the right panel.

The heavy lines in the left panel represent a situation where hits have a value of 1.5, misses -1.5 , false alarms -0.7 , and correct rejections 0.7. The optimum criterion in this case (right panel, magenta) shifts to a smaller (more liberal) value. This results in more hits and fewer misses at the cost of more false alarms and fewer correct rejections, reflecting the relative value of these outcomes. The optimum criterion is the value of y that satisfies the following equation:

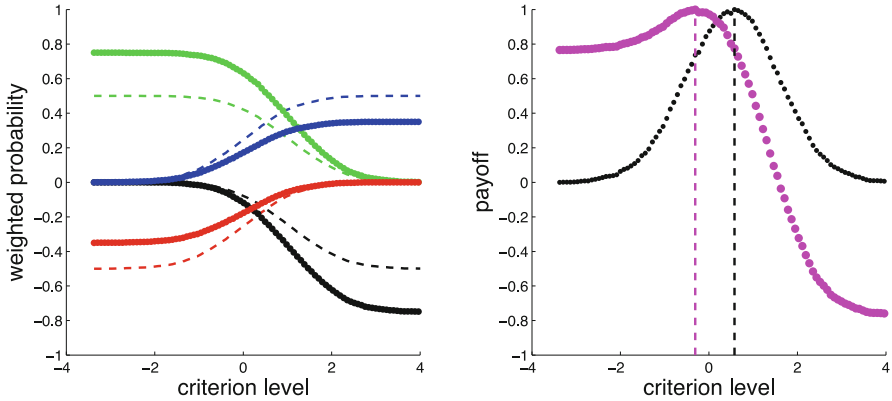


Fig. 5.6 Effect of payoffs on optimal decision criteria. (*Left*) All outcomes have same absolute value (*dashed lined*) or different values (*solid lines*). (*Right*) Total payoff vs. criterion for equal outcome value (*black*) and unequal value (*magenta*)

$$(V_h - V_m) * p(S) * f_s(y) = (V_{cr} - V_{fa}) * p(N) * f_n(y)$$

where $V_x(x = h, m, fa, cr)$ is the value of a hit, miss, false alarm, or correct rejection. Again, V_m and V_{fa} usually have negative values so that when all outcomes have equal weight (0.5), the terms $(V_h - V_m)$ and $(V_{cr} - V_{fa})$ can be replaced by 1.0.

The values assigned to different outcome classes may reflect economic value, such as subjective utility. They may also reflect emotional value (intensity) and valence (positive or negative). Stimuli that are associated with high outcome values may automatically attract attention, regardless of whether the outcome is positive or negative [15].

5.4 Effects of Pooling Over Multiple Detectors

The above considerations apply to the case of a single detector (e.g., a neuron) but can be readily extended to multiple detectors. In the simplest case, all detectors have the same inputs, sensitivity, and noise characteristics. All observations therefore have equal weight. One observation from each of two detectors is the same as two observations from one detector. However, even this simple case presents an opportunity to test different rules for pooling across detectors. Furthermore, we can examine how detectability improves with the number of detectors and observe the effects of correlations among detector responses. Studies in monkeys have found that the spike count correlation between nearby visual cortical neurons is roughly 0.1–0.2 [16, 17] and that these correlations are reduced by attention [18, 19].

To analyze the activity in multiple detectors, it is useful to first build an activity matrix. Each row in this matrix represents an individual detector (neuron), and each

Table 5.3 Activity matrix with each cell $y(i,j)$ representing an average firing rate for neuron i on trial j

	Trial 1	Trial 2	Trial 3	...
Neuron 1	$y(1,1)$	$y(1,2)$	$y(1,3)$...
Neuron 2	$y(2,1)$	$y(2,2)$	$y(2,3)$...
Neuron 3	$y(3,1)$	$y(3,2)$	$y(3,3)$...
...

column is a single observation period (trial). The values in each cell thus represent the activity of a single detector on a single trial (Table 5.3):

From this activity matrix, we can construct joint ROC functions in several ways. The simplest method is to treat each cell, $y(i,j)$, as an independent observation. The result is that the presence of multiple detectors increases the number of observations at a given time, but otherwise confers no improvement in detectability. In other words, if all neurons are equal, then adding neurons does not change the joint ROC curve. It is the same as simply gathering additional observations from a single neuron.

To gain any advantage from multiple neurons, the responses must be aggregated in some manner. One method is simply to average observations prior to constructing the ROC function. One can either average over trials for each neuron or over neurons within each trial. Both methods reduce the number of effective observations underlying the ROC, but have the advantage that the variability of those observations may be substantially reduced. Figure 5.7 shows an example with 12 neurons. In the left panel are the responses of a representative pair of neurons collected over many trials (blue dots = stimulus absent, red dots = stimulus present). The trial-to-trial responses are weakly correlated (Pearson’s correlation coefficient = 0.2). The right panel shows the ROC function for each of the 12 neurons individually (blue lines) and for the ensemble (red dots) when activity in each trial is averaged across neurons before computing the ROC. The average ROC area for each neuron alone is 0.76, while the joint ROC has area = 0.86. In this example, all neurons have the same sensitivity, and their contributions are weighted equally. One could alternatively construct a weighted average such that the contribution of each neuron would be weighted by its reliability, for example, by dividing by the standard deviation or variance of the spike count distribution. This would result in a more Bayesian style of combining responses.

Figure 5.8 (left panel) shows how detectability increases with the number of neurons. When the response of each neuron is independent of the other neurons, detectability starts to saturate when there are about 32 neurons in the pool (black curve). The exact number of neurons at which saturation occurs is not fixed, but typically depends on the characteristics of the signal and noise, as well as the method of pooling [20].

The advantage of pooling responses across neurons is reduced when their activity is correlated. For example, if the degree of correlation for every pair of cells in the population is $r = 0.2$ (as depicted in Fig. 5.7, left panel), then the area under the joint ROC is represented by the red curve in Fig. 5.8 (left panel). Here, the optimum detectability reaches only 85 % of that obtained when there is no

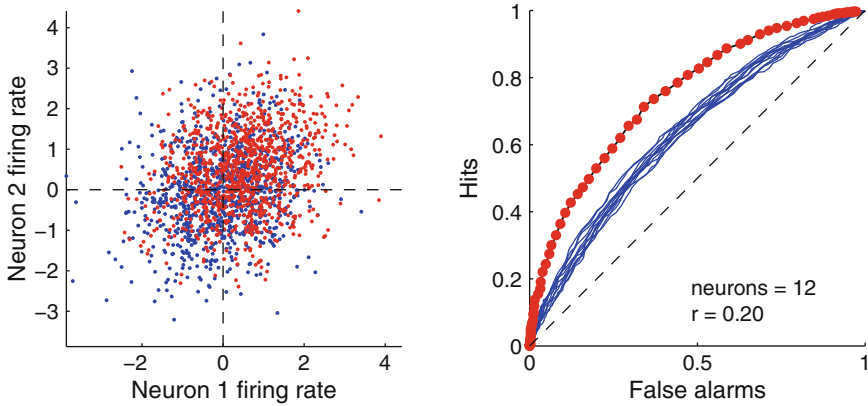


Fig. 5.7 (Left) Firing rates of two simulated neurons whose trial-to-trial responses are weakly correlated. Blue dots are for stimulus-absent, and red dots are for stimulus-present trials. (Right) ROC curves for 12 weakly correlated neurons. Blue lines are ROCs computed for each neuron individually. Red dots are the ROC curve when responses are pooled across all neurons

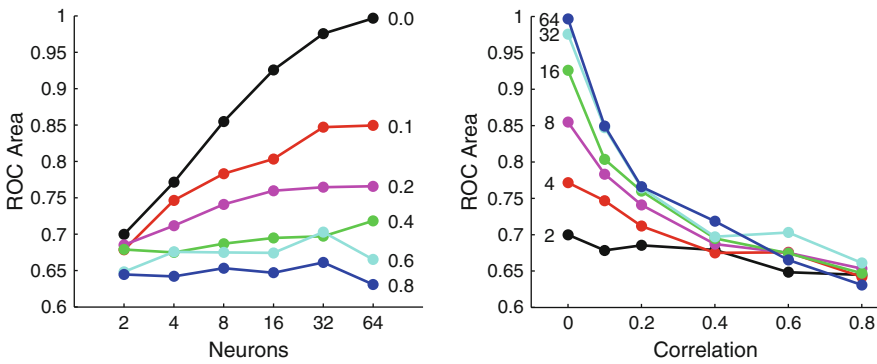


Fig. 5.8 Impact of neuronal pool size (left panel) and between-neuron response correlations (right panel) on detectability

correlation between neurons. When the correlation is 0.6 (green curve), there is almost no advantage of pooling. The same data are replotted as a function of correlation strength (Fig. 5.8, right panel). The curves for different pool sizes all come together between $r=0.4$ and $r=0.6$, indicating that the strength of correlation that eliminates the advantage of pooling in this case is about 0.5.

If positive correlations among neurons reduce the benefits of pooling, then it might be expected that negative correlations would have the opposite effect. Figure 5.9 shows an example where the pool size is 2 neurons and the trial-to-trial correlation in firing rate is -0.9 . The negative correlation reduces the overlap of signal and noise distributions. The resulting joint ROC has an area of 0.95. This can be compared to the case where the correlation is 0 and the joint ROC area is

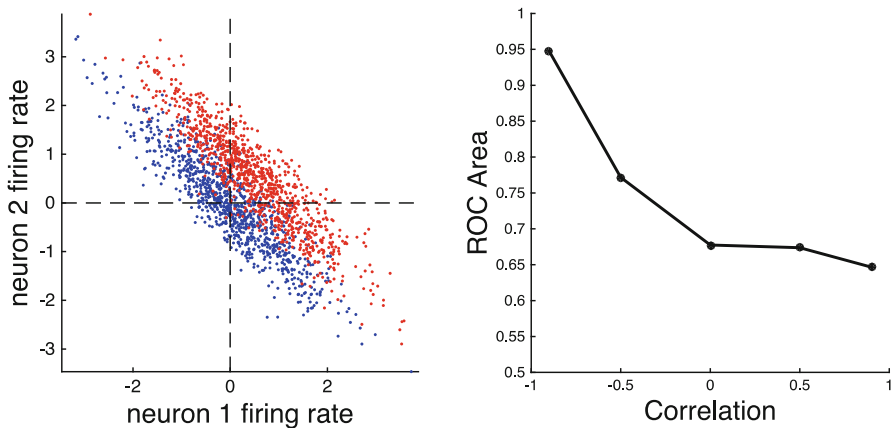


Fig. 5.9 Impact of negative correlations on detectability for a pool of 2 neurons. (Right) signal (red) and noise (blue) distributions. (Left) ROC area as a function of correlation coefficient

0.69. It would therefore seem that negative correlations are capable of producing huge improvements in detectability. However, there are some important caveats. Foremost among these is that for a pool size greater than 2, it is impossible for all the pairwise correlations to be negative. If neurons 1 and 2 are negatively correlated, and 2 and 3 are also negatively correlated, then 1 and 3 must be positively correlated. In other words, the correlation matrix must be positive semi-definite. As long as this constraint is satisfied, a pool of neurons with some negative correlations might provide significantly better detectability than a pool of independent, uncorrelated neurons.

For weaker stimuli, the advantages of increasing the size of the neuronal pool are greater; however, the effects of positively correlated activity are more devastating. For example, when there is no correlation between neurons, the improvement in detectability of a weak stimulus may not start to saturate until the pool size reaches over 1000 neurons. However, even a weak ($r = 0.1$) correlation can eliminate the advantage of pooling altogether.

Another method for pooling across neurons is to compute the probability of hits and false alarms for each neuron individually and then sum the probabilities across neurons. The probabilistically summed hit and false alarm rates can then be used to compute the joint ROC. The pooled probability of a hit and false alarm is as follows:

$$p(\text{hit}) = 1.0 - [p(\text{miss}_1) \& p(\text{miss}_2) \& \dots \& p(\text{miss}_n)]$$

$$p(\text{fa}) = 1.0 - [p(\text{cr}_1) \& p(\text{cr}_2) \& \dots \& p(\text{cr}_n)]$$

where the subscripts (1 ... n) index the individual neurons and n is the total number of neurons. The probability of a miss or correct rejection is

calculated across all trials for each neuron. These calculations are done for each criterion level to construct the joint ROC curve. Note that because the firing rates may be correlated across neurons, the probabilities of misses or correct rejections are not independent. Thus, the right sides of these equations must be calculated using the formula for the joint probability of dependent events:

$$p(a \& b \& c \& d \& \dots) = p(a) * p(b|a) * p(c|a \& b) * p(d|a \& b \& c) \dots$$

These results suggest that attention can improve signal detection by reducing correlated activity among neurons or by selecting the responses of neurons whose activity is maximally uncorrelated (or negatively correlated). However, we have so far only dealt with the case of detectors with uniform sensitivity. In general, neurons may have different sensitivities (responsiveness) and baseline firing rates. Furthermore, the degree of correlated activity between pairs of neurons is likely to vary across the population, rather than being constant for all pairs as in the simulations presented here.

The above considerations apply to the case where all neurons in the pool respond to the signal. In this case, increasing the number of neurons in the pool increases stimulus detectability, though the marginal improvement may at times be small. However, it is unusual for all detectors to be sensitive to the stimulus. In general, a given stimulus will be represented by a small fraction of the relevant population of neurons, i.e., those whose front-end filtering properties (e.g., spatial and feature selectivity) are appropriately matched to the stimulus. We can call this the “signal pool.” The rest of the neurons in the brain (the “noise pool”) contribute nothing to detection of the stimulus. In fact, their activity is deleterious to performance as it represents background noise. One of the great problems of attention is how to select responses from the signal pool while ignoring or suppressing activity in the noise pool. The problem is compounded by the fact that individual neurons can switch from one pool to the other at any given time, depending on the stimuli present in the environment and the organism’s behavioral goals.

5.5 Uncertainty and Cueing Effects

One of the most common behavioral paradigms for studying attention is to provide observers with prior information (a cue) about a target whose properties are uncertain [4, 21, 22]. For example, observers might be asked to detect a low contrast target presented at a location that is randomized from trial-to-trial, thus introducing spatial uncertainty. At some time before the target appears, a high contrast cue is presented at a location that is more or less predictive of the target location. Such cues often improve performance accuracy, but whether these improvements are due to enhanced stimulus detectability or reduction in decision uncertainty has been subject to much debate [23–27].

Figure 5.10 illustrates trial conditions from a task where the stimulus (a vertically oriented grating patch) can appear at one of two locations (essentially the same task as used by [21]). The subject's task is to report the presence of the target. On all trials, the cue is equally likely to occur at either location. On half the trials, there is no stimulus ("catch" trials). On the other trials, the stimulus is preceded by a cue that predicts where the stimulus is likely to appear (i.e., it may or may not be a "valid" cue). The predictiveness of the cue is referred to as its "validity." If the cue is 80 % valid, then the stimulus appears at the cued location on 80 % of the trials and at the uncued location on 20 % of the trials. If the cue is valid on 80 % of the signal-present trials and invalid on 20 %, then the signal probability is 0.4 at the cued location and 0.1 at the uncued location. The cue not only attracts attention but allows the observer to use a more liberal criterion for responding that the stimulus is present. This should result in a higher percent correct on cued vs. uncued trials. It may also lead to shorter reaction times.

Consider the responses of two detectors, one at the cued location and another at the uncued location. What level of performance can be achieved by combining responses from the two detectors with equal weight? Each detector experiences a signal probability of 0.25, because the stimulus is present on half the trials and its location is randomized. Figure 5.11 (top left) illustrates the theoretical percent correct (hits + correct rejections divided by total trials) for detecting the stimulus as a function of criterion level for both detectors when the cue validity is 50 % (i.e., the

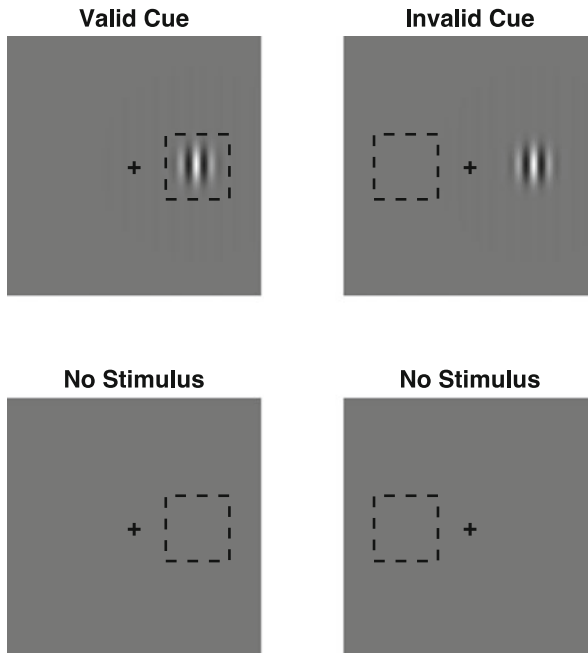


Fig. 5.10 Spatial cueing task with vertically oriented target

cue location is uncorrelated with stimulus location). Because the signal probability at each detector is 0.25, the optimal criterion (vertical dashed lines) is relatively conservative and is the same for both detectors.

The same calculations are shown for the case where cue validity is 80 % (Fig. 5.11, top right). Here, the optimal criterion is more liberal for the detector at the cue location (blue) because the signal probability (0.4) is higher at that location. The stimulus probability at the uncued location is only 0.1. This leads to the counterintuitive observation that the uncued detector can actually achieve better performance than the cued detector. This happens because an observer can use a very conservative criterion for the uncued detector. In fact, they can say “no” (i.e., reject the hypothesis that the stimulus is present at the detector location) on every trial and be correct 90 % of the time, regardless of the state of the detector.

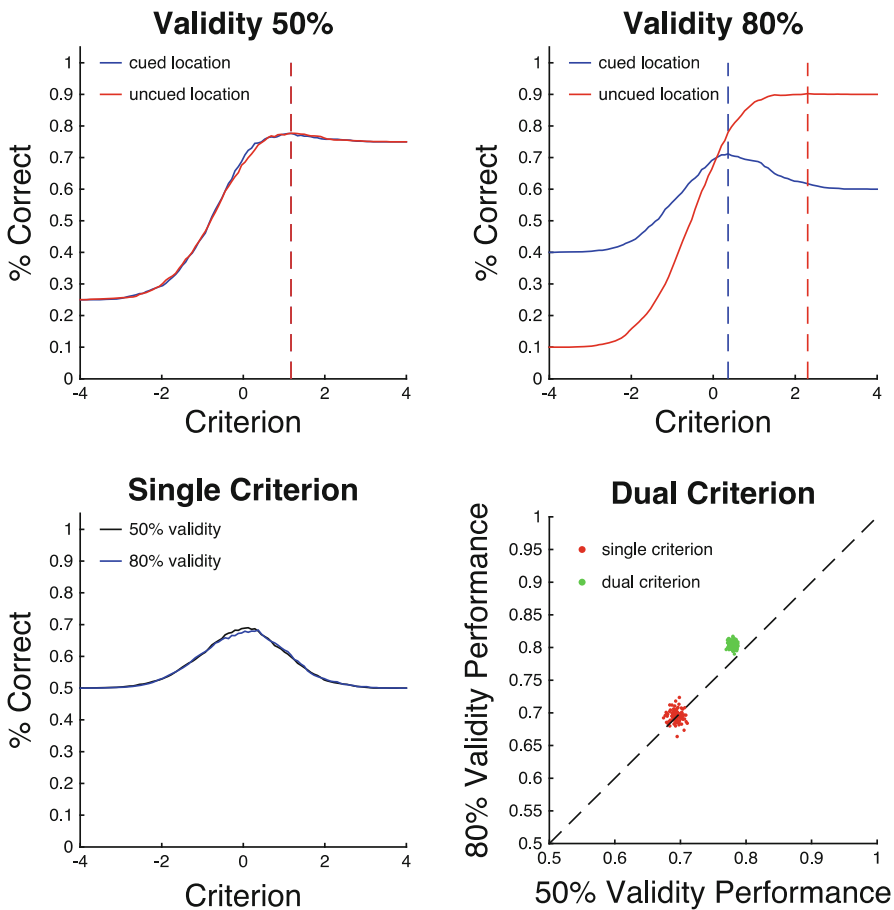


Fig. 5.11 Effects of cueing analyzed by SDT

It may appear that, in the case of 80 % cue validity, one should be able to achieve high performance simply by using the response of the better detector. Unfortunately, each detector only provides partial information (whether or not the stimulus is present at the detector location). Both detector responses must be combined to determine the overall response of the observer. Overall performance is calculated by adding up hits and correct rejections from all detectors and dividing by the total number of observations (number of trials x number of detectors). Surprisingly, if one is limited to using the same criterion for both detectors, then valid cues offer no advantage. The percent correct is the same for both 50 % and 80 % validity (Fig. 5.11, bottom left), and this is true regardless of criterion. In a sense, this is understandable as the cues provide information only about likely stimulus location, whereas the observer's job is to report stimulus presence.

One way that valid cues can yield an advantage is if the observer is allowed to choose the optimum criterion independently for each detector. Figure 5.11 (bottom right) shows performance for the case where the optimum criterion for each detector is used (green). This is compared to the case of a single criterion that optimizes performance for both detectors (red). The advantage of valid cues is small. This is partly due to the fact that the proportion of catch trials is large (50 %). Reducing the proportion of catch trials increases the performance advantage provided by valid cues. Whether or not subjects are capable of maintaining multiple decision criteria at the same time is an open question [28].

Other approaches to understanding cueing effects have been suggested. Cueing effects can be modeled using Bayesian statistics, which leads to similar conclusions about the advantages of valid cues [7]. In all of the above, the assumption is that valid cues affect the decision process but not signal quality. If valid cues enhance signal strength, then that advantage would add to the advantage one can achieve by adjusting decision criteria.

5.6 Signal Detection Over Time

Attention can increase the rate of information processing [29]. Hence, models are needed that account for both improvements in detectability and response time. However, the preceding discussion applies only to signals that are non-time-varying in the sense that they remain constant over the duration of a given observation period or "trial." The assumption is that, on each trial, the observer draws a single sample from the distribution of internal states of each detector and a decision is made based on those samples. These models can predict performance accuracy, but not the amount of time needed to respond at a given level of accuracy. Adding the dimension of time allows observers to draw multiple samples from each detector and to integrate the evidence provided by those samples, before reaching a decision.

In the 1940s, Wald [30], and others, developed the theory of sequential sampling as a way to calculate the incremental evidence provided by each sample and, thus, how many samples are needed for a given level of performance. If samples are

drawn at a steady rate, this number corresponds to response time. Specifically, Wald developed the *sequential probability ratio test* (SPRT), which integrates the incremental information provided by each sample and also specifies a stopping rule, i.e., the amount of integrated evidence needed to achieve a given level of accuracy, defined in terms of percentage correct (hits and correct rejections) or incorrect (false alarms and misses). This test is derived from the Neyman-Pearson lemma which states that the likelihood ratio test maximizes the probability of detection for a given probability of false alarms [31].

The problem addressed by the SPRT is how to quantify the information in each sample so that it can be combined with other samples. The optimal way to do this is to start with the likelihood that a given sample, y , was drawn from the signal-present or signal-absent probability densities, i.e., $p(y|S)$ and $p(y|N)$. The next step is to compute the log of the likelihood ratio: $x = \log[p(y|S)/p(y|N)]$. The quantity, x , represents the momentary evidence favoring hypothesis H1: signal present vs. H0: signal absent. The process is iterated by repeatedly drawing samples, calculating the log of the likelihood ratio, and adding that incremental evidence to the total evidence accumulated from previous samples:

$$x_{t+1} = x_t + \log \left[\frac{p(y_{t+1}|S)}{p(y_{t+1}|N)} \right].$$

The accumulation of evidence continues until x reaches a threshold value, or boundary. There are two boundaries: if x first reaches bound A , H1 is accepted (e.g., observer responds “yes”), and if x reaches bound B , H1 is rejected (response is “no”). The values of A and B are calculated to yield a predetermined level of performance accuracy. If α is the desired false alarm rate and β the desired miss rate, then $A = \log[(1 - \beta)/\alpha]$, and $B = \log[\beta/(1 - \alpha)]$. The bounds can also be calculated in terms of the hit and correct rejection rates, as hit rate = $1 - \beta$ and correction rejection rate = $1 - \alpha$.

The SPRT can be thought of as a one-dimensional diffusion-to-bound process [32, 33], wherein the decision variable, x , takes a random walk that starts at zero and ends at one of the two bounds. This can be written as $dx/dt = r + u(0,s)$, where r is the mean drift rate and u is the momentary noise represented by a random variable drawn from some distribution, typically a Gaussian with mean = 0 and standard deviation = s . The random element guarantees that, given enough time, x will hit one bound or the other even if the drift rate is zero. The diffusion parameters (r, s) as well as the bounds (A, B) can be fit to experimental data for accuracy and reaction time [34].

Figure 5.12 shows simulations based on the SPRT where the likelihood density functions are Gaussians. The outcomes can be classified as hits (blue) and correct rejections (red), as well as misses and false alarms (not shown). The proportions of correct and incorrect trials as well as the response time distributions for each class of outcome are fully determined by the log-likelihood ratio and the boundaries (Fig. 5.12, below).

The standard SDT notions of detectability and response bias are built into the SPRT. Detectability depends on the rate of evidence accumulation, drift rate, and the variance in drift rate or momentary noise. Response bias occurs when the bounds are

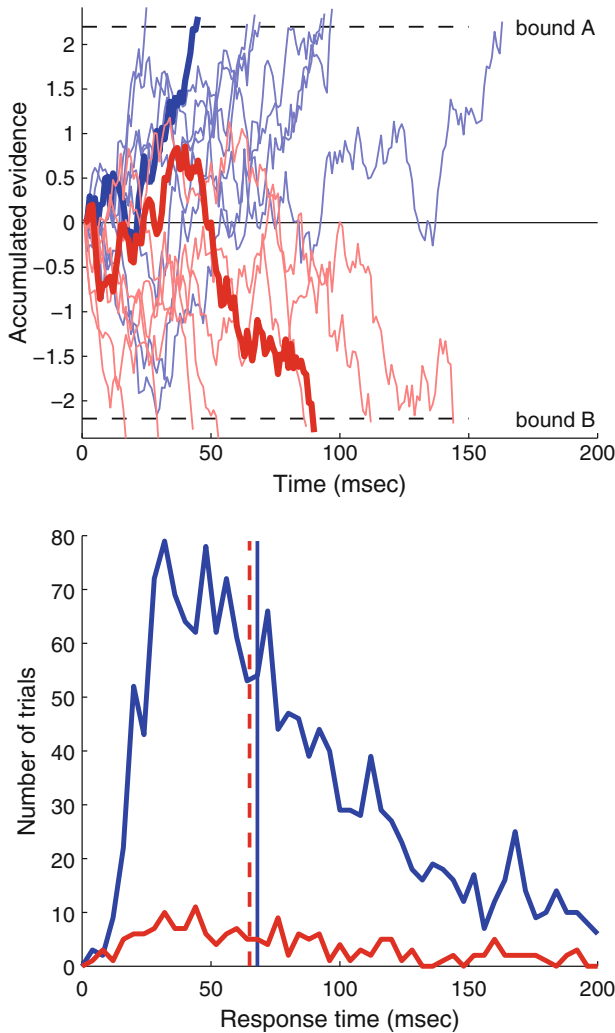


Fig. 5.12 Simulations of SPRT. (Top) Blue lines represent stimulus-present trials. Red are stimulus-absent trials. (Bottom) RT distributions for stimulus-present (blue) and stimulus-absent (red) trials

asymmetric so that the diffusion process starts from a position closer to one bound than the other. The SPRT effectively has two independent criteria, whereas the static SDT model has only one. In the SPRT, the outcome classes are more independent. For example, it is possible to maintain a constant hit rate while varying the false alarm rate. Thus, the trade-off between hits and false alarms that is characteristic of the static SDT model does not hold for the SPRT.

If attention increases signal quality by reducing signal to noise, the effect on the SPRT will be to increase the rate of evidence accumulation [35]. This is equivalent to

improving stimulus detectability. Others have incorporated salience and economic value by modulating drift rate [36].

The SPRT also provides a solution to the problem of pooling responses across multiple detectors. The summation of log-likelihoods applies not only to the integration of multiple samples from a single detector, but also to the integration of individual samples from multiple detectors. Given a set of samples from multiple detectors, one can simply sum the log of the likelihood ratios to obtain an estimate of the evidence that a stimulus is present. This could be called the parallel probability ratio test (PPRT). This calculation can be performed at every moment in time. The conversion from raw samples to log-likelihoods takes into account the signal-to-noise of each detector and thus provides a common metric for integrating responses from detectors with different sensitivities, filtering properties, and noise characteristics.

Computing the SPRT in parallel across the visual scene using a 2D array of detectors results in a detectability salience map. This is illustrated in Fig. 5.13 with a 40×40 array of detectors. The signal can occur at one of two locations (lower left or upper right). Detectability (hit rate – false alarm rate) for each detector is plotted (black indicates detectors with high false alarms, white indicates detectors with high hit rate). The cueing paradigm described in Fig. 5.10 was implemented with two different cue validities. Cue validity is implemented by biasing the starting point of the decision processes at the cued and uncued locations [37]. When the cue validity is 0.5, the cue provides no information, the signal and noise distributions have equal area, and detectability is equal at the two locations. When the cue validity is 0.55, the signal appears at the cued location 55 % of the time and at the uncued location 45 % of the time. This enhances the detectability at the cued location and reduces detectability at the uncued location. Figure 5.13 plots stimulus detectability, but SPRT-computed salience can also be expressed in terms of response time. If detectability and response time are combined, it is possible to calculate the information processed by the observer in terms of bits/second

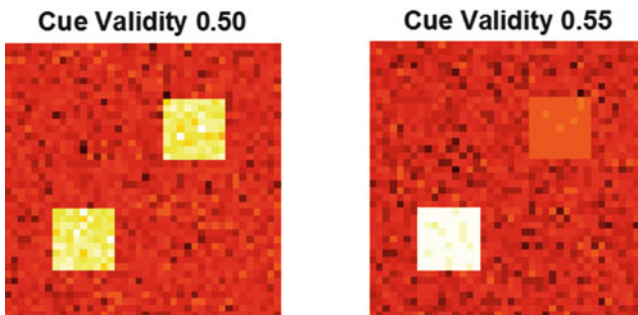


Fig. 5.13 Salience maps computed using SPRT. Both maps show the detectability of a signal that can occur at one of two locations. (*Left*) signal occurs at either location with equal probability. (*Right*) signal occurs at cued location (*lower left*) 55 % of the time and uncued location (*upper right*) 45 % of the time. Intensity indicates stimulus detectability

(e.g., [38]). Furthermore, the SPRT allows the observer to adjust the decision boundaries at the cued and uncued location, which should also affect the relative salience.

5.7 Conclusion

Signal detection theory provides a simple yet powerful framework for understanding how observers respond to weak signals in the environment. The theory makes a clear distinction between detection and response selection. Attention can improve signal detection by increasing the gain of sensory responses while reducing noise. For a fixed level of detectability, attention can further improve performance by optimizing decision criteria. When there are multiple detectors, attention can improve detectability by de-correlating responses and by selectively monitoring detectors that are more sensitive to the stimulus by virtue of their receptive field location, feature selectivity, or other properties.

Acknowledgements This work was supported by NIH grant MH059244. We thank Prof. Norma Graham and Dr. Greg Jensen for a critical reading of the manuscript.

References

1. James, W. (1890). *The principle of psychology*. New York: Henry Holt & Co.
2. Broadbent, D. (1958). *Perception and communication*. London: Pergamon Press.
3. Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review*, 75, 522–536.
4. Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology. General*, 109, 160–174.
5. Eriksen, C., & St. James, J. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, 40(4), 225–240.
6. Graham, N. V. S. (1989). *Visual pattern analyzers*. New York: Oxford University Press.
7. Eckstein, M. P., Peterson, M. F., Pham, B. T., & Droll, J. A. (2009). Statistical decision theory to related neurons to behavior in the study of covert visual attention. *Vision Research*, 49, 1097–1128.
8. Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, 70(1), 80–90.
9. Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18, 193–222.
10. Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
11. Mcmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). New York: Lawrence Erlbaum Associates, Psychology Press.
12. Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, 435, 439.

13. Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., Britton, C. A., Brown, M. L., Davis, P. L., Ferris, J. V., Fuhrman, C., Golla, S. K., Katyal, S., Lacomis, J. M., McCook, B. M., Thaete, F. L., & Warfel, T. E. (2003). Prevalence effect in a laboratory environment. *Radiology*, 228, 1–14.
14. Wickens, T. D. (2001). *Elementary signal detection theory*. New York: Oxford University Press.
15. Maunsell, J. H. (2004). Neuronal representations of cognitive state: Reward or attention? *Trends in Cognitive Sciences*, 8(6), 261–265.
16. Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, 16(4), 1486–1510.
17. Zohary, E., Shadlen, M. N., & Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485), 140–143.
18. Cohen, M. R., & Maunsell, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, 12(12), 1594–1600.
19. Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2009). Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron*, 63(6), 879–888.
20. Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, 16(4), 1486–1510.
21. Bashinski, H. S., & Bacharach, V. R. (1980). Enhancements of perceptual sensitivity as the result of selectively attending to spatial locations. *Perception & Psychophysics*, 28, 241–248.
22. Mertens, J. J. (1956). Influence of knowledge of target location upon the probability of observation of peripherally observable test flashes. *Journal of the Optical Society of America*, 46(12), 1069–1070.
23. Downing, C. J. (1988). Expectancy and visual-spatial attention: Effects on perceptual quality. *Journal of Experimental Psychology: Human Perception and Performance*, 14(2), 188–202.
24. Hawkins, H. L., Hillyard, S. A., Luck, S. J., Mouloua, M., Downing, C. J., & Woodward, D. P. (1990). Visual attention modulates signal detectability. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4), 802–811.
25. Muller, H. J., & Findlay, J. M. (1987). Sensitivity and criterion effects in the spatial cueing of visual attention. *Perception & Psychophysics*, 42, 383–399.
26. Shaw, M. L. (1984). Division of attention among spatial locations: A fundamental difference between detection of letters and detection of luminance increments. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X* (pp. 109–121). Hillsdale: Erlbaum.
27. Müller, H. J., & Findlay, J. M. (1987). Sensitivity and criterion effects in the spatial cuing of visual attention. *Perception & Psychophysics*, 42(4), 383–399.
28. Gorea, A., & Sagi, D. (2005). Decision and attention. In L. Itti, G. Rees, & J. Tsotsos (Eds.), *Neurobiology of attention*. New York: Elsevier.
29. Carrasco, M., & McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences*, 98(9), 5363–5367.
30. Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2), 117–186.
31. Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231(694–706), 289–337.
32. Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
33. Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, 116(2), 283–317.
34. Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review*, 9(3), 438–481.

35. Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *PNAS*, *108*, 13852–13857.
36. Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *PNAS*, *110*(40), E3858–E3867.
37. Sperling, G. (1984). A unified theory of attention and signal detection. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 103–181). London: Academic Press.
38. Teichert, T., Ferrera, V. P., & Grinband, J. (2014). Humans optimize decision-making by delaying decision onset. *PLoS One*, *9*(3), e89638.