Castleman
Digital Image Processing

Chapter 11

# FILTER DESIGN

## INTRODUCTION

In Chapters 9 and 10, we laid the groundwork for the analysis of linear filtering operations. Linear filters can be implemented either by convolution in the time (or spatial) domain or by multiplication in the frequency domain. In later chapters, we deal with the techniques for and limitations of implementing linear filtering digitally. In this chapter, however, we discuss techniques for designing filters to accomplish particular goals. To gain insight, we shall first examine the time domain and frequency domain behavior of certain simple but useful filters. Later in this chapter, we approach the problem of designing filters that are optimal for doing a specific job.

As in Chapters 9 and 10, we shall perform the analysis with one-dimensional (time) signals for simplicity of mathematics. The generalization to two dimensions is straightforward.

## EXAMPLES OF COMMON FILTERS

In this section, we consider some conceptually simple filters in order to gain insight into the time domain and frequency domain characteristics of filters and their effect upon input signals.





### The Ideal Bandpass Filter

Suppose we desire to implement, by convolution, a filter that passes energy only at frequencies between $f_1$ and $f_2$, where $f_2 > f_1$. The desired transfer function is given by

$$G(s) = \begin{cases} 1 & f_1 \le |s| \le f_2 \\ 0 & \text{elsewhere} \end{cases} \tag{1}$$

and shown in Figure 11-1. Since $G(s)$ is an even rectangular pulse pair, it can be thought of as a rectangular pulse convolved with an even impulse pair. If we let

$$s_0 = \tfrac{1}{2}(f_1 + f_2) \quad \text{and} \quad \Delta s = f_2 - f_1 \tag{2}$$

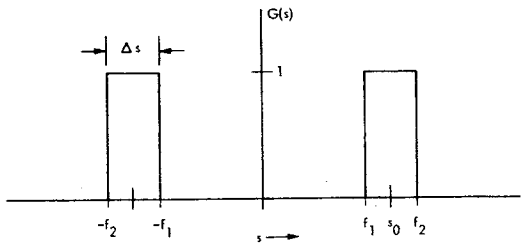



**Figure 11-1** Ideal bandpass transfer function

we can write the transfer function of the ideal bandpass filter as

$$G(s) = \Pi\left(\frac{s}{\Delta s}\right) * [\delta(s - s_0) + \delta(s + s_0)] \tag{3}$$

With the transfer function expressed in this form, we can easily write the impulse response

$$g(t) = \Delta s \frac{\sin(\pi \Delta s t)}{\pi \Delta s t} 2\cos(2\pi s_0 t) = 2\Delta s \frac{\sin(\pi \Delta s t)}{\pi \Delta s t} \cos(2\pi s_0 t) \tag{4}$$

Since $\Delta s < s_0$, Eq. (4) describes a cosine of frequency $s_0$ enclosed in a sin $(x)/x$ envelope having frequency $\Delta s/2$. This impulse response is graphed in Figure 11-2. The number of cosine cycles between envelope zero crossings depends on the relationship between $s_0$ and $\Delta s$. Notice that if $s_0$ is held constant and $\Delta s$ becomes small, the envelope expands to include more and more cosine cycles between zero crossings. As $\Delta s$ approaches zero, the impulse response approaches a cosine. In the limiting case, the convolution actually becomes a cross-correlation of the input with the cosine at frequency $s_0$.

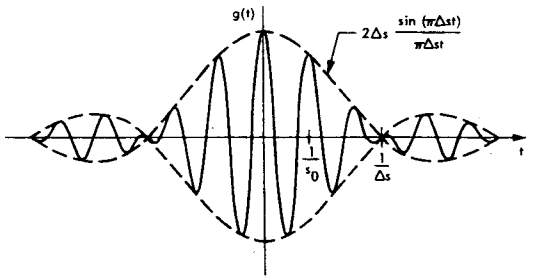



**Figure 11-2** Ideal bandpass impulse response

### The Ideal Bandstop Filter

Suppose now we desire a filter that passes energy at all frequencies except for a band between $f_1$ and $f_2$, where $f_2 > f_1$. This transfer function is given by

$$H(s) = \begin{cases} 0 & f_1 \le |s| \le f_2 \\ 1 & \text{elsewhere} \end{cases} \tag{5}$$

and graphed in Figure 11-3. For convenience, we again let $s_0$ be the center frequency

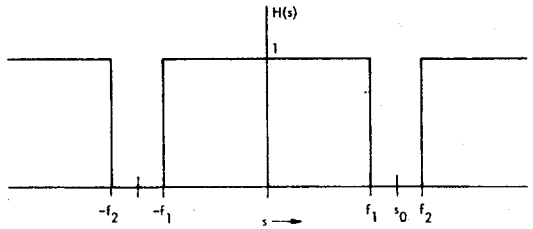



**Figure 11-3** Ideal bandstop transfer function

and $\Delta s$ the bandwidth (Eq. 2). Now we can write the transfer function as one minus a bandpass filter

$$H(s) = 1 - \Pi\left(\frac{s}{\Delta s}\right) * [\delta(s - s_0) + \delta(s + s_0)] \tag{6}$$

from which the impulse response is

$$h(t) = \delta(t) - 2\Delta s \frac{\sin(\pi\Delta st)}{\pi\Delta st} \cos(2\pi s_0 t) \tag{7}$$

The impulse response is graphed in Figure 11-4. Its behavior with changing bandwidth



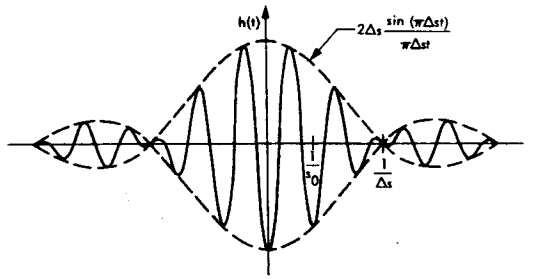



**Figure 11-4** Ideal bandstop impulse response

and center frequency is similar to that of the bandpass filter, which it resembles. If $\Delta s$ is small, this filter is referred to as a "notch filter."

### The General Bandpass Filter

We now consider a class of bandpass filters constructed in the following way. We select a nonnegative unimodal function $F(s)$ and convolve it with an even impulse pair at frequency $s_0$. This yields a bandpass transfer function, as shown in Figure 11-5.

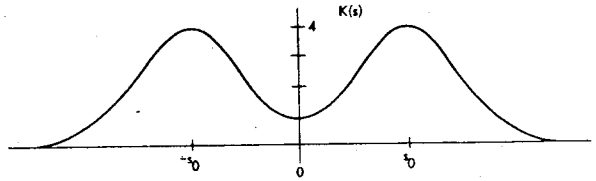



**Figure 11-5** The general bandpass filter

The transfer function is given by

$$K(s) = F(s) * [\delta(s - s_0) + \delta(s + s_0)] \tag{8}$$

and the impulse response by

$$k(t) = 2f(t) \cos(2\pi s_0 t) \tag{9}$$

This impulse response is a cosine of frequency $s_0$ in an envelope that is the inverse Fourier transform of $F(s)$.

Suppose, for example, that $F(s)$ is a Gaussian

$$K(s) = Ae^{-s^2/2\sigma^2} * [\delta(s - s_0) + \delta(s + s_0)] \tag{10}$$

Then the impulse response becomes

$$k(t) = \frac{2A}{\sqrt{2\pi\sigma^2}} e^{-t^2/2\sigma^2} \cos(2\pi s_0 t) \tag{11}$$

This impulse response is a cosine in a Gaussian envelope. It is graphed in Figure 11-6. Notice that we could easily generate a class of bandstop filters as before.

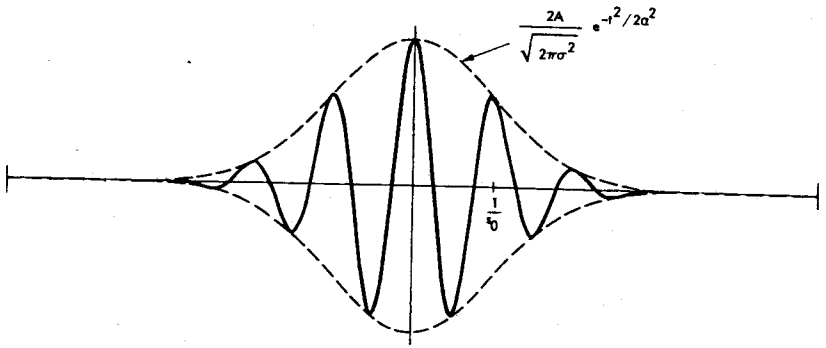



**Figure 11-6** The Gaussian bandpass filter

### The Gaussian High-Frequency Enhancement Filter

The term *high-frequency enhancement filter* is generally taken to describe a transfer function that is unity at zero frequency and increases away from the origin. A high-frequency enhancement filter may either level off at some value greater than one or, more commonly, may fall back toward zero at high frequencies. In the latter case, the high-frequency enhancement filter is merely a type of bandpass filter with the restriction of unity gain at zero frequency. In practice, it is often desired to have less than unity gain at zero frequency to reduce the contrast of large, slowly varying components in the image.

We can produce a high-frequency enhancement transfer function by expressing it as the difference of two Gaussians of different widths.

$$H(s) = Ae^{-s^2/2\alpha_1^2} - Be^{-s^2/2\alpha_2^2} \qquad A \geq B, \alpha_1 > \alpha_2$$

This is shown in Figure 11-7. The impulse response of this filter is given by

$$h(t) = \frac{A}{\sqrt{2\pi\sigma_1^2}} e^{-t^2/2\sigma_1^2} - \frac{B}{\sqrt{2\pi\sigma_2^2}} e^{-t^2/2\sigma_2^2} \qquad \sigma_i = \frac{1}{2\pi\alpha_i} \tag{12}$$

and graphed in Figure 11-8. Notice that the broad Gaussian in the frequency domain produces a narrow Gaussian in the time domain and vice versa. The impulse response



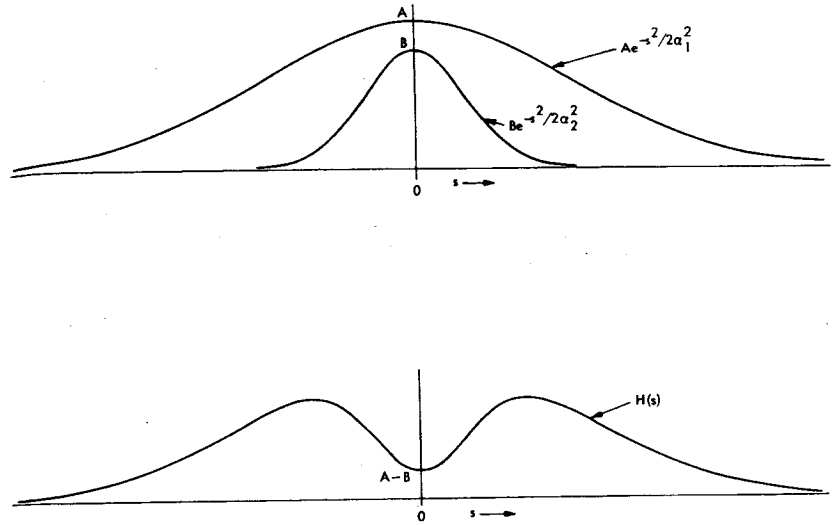



**Figure 11-7** The Gaussian high-frequency enhancement transfer function

shown in Figure 11-8 is typical of bandpass and high-frequency enhancement filters, having a positive pulse situated in a negative dish.

If we let $\alpha_1$ approach infinity, the narrow Gaussian in the time domain narrows down to an impulse and the filter has the form shown in Figure 11-9. Notice that the difference between a filter that rolls off (returns toward zero) at high frequencies and one that does not is the width of the central pulse in the time domain.

### Rules of Thumb for High-Frequency Enhancement Filter Design

In this section, we develop two approximate rules to estimate the behavior of high-frequency enhancement filters. Suppose the impulse response of the filter is expressed as a narrow pulse minus a broad pulse,

$$h(t) = h_1(t) - h_2(t) \tag{13}$$

as illustrated in Figure 11-10. We know that the transfer function $H(s)$ will have the general shape of a high-frequency enhancement filter. We would like to estimate the transfer function at zero frequency to determine its effect on the contrast of large objects within the image. We also would like to estimate the maximum value the transfer function takes on at any frequency.

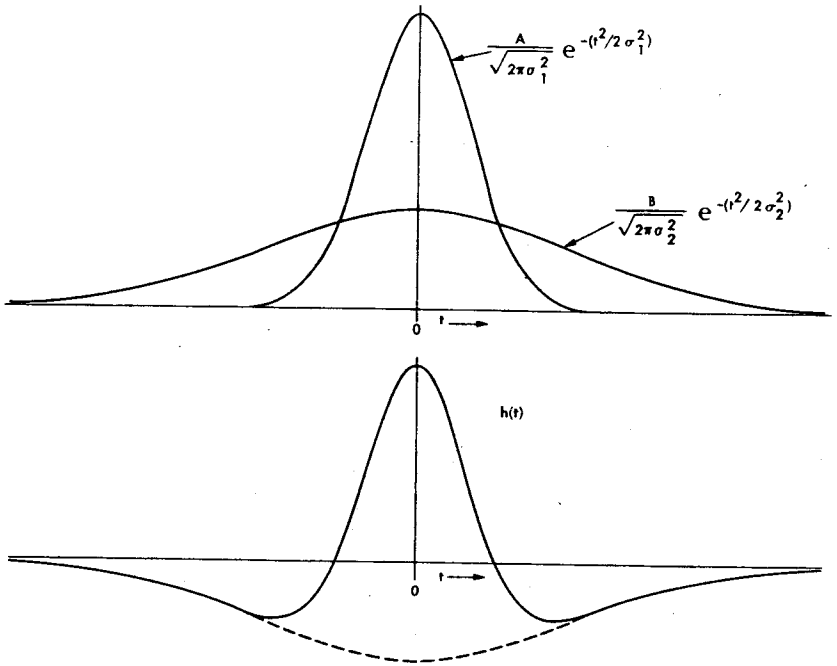



**Figure 11-8** The Gaussian high-frequency enhancement impulse response

If we write the Fourier transform of Eq. (13) and substitute the value $s = 0$, we obtain

$$H(0) = \int_{-\infty}^{\infty} h(t)\,dt = \int_{-\infty}^{\infty} h_1(t)\,dt - \int_{-\infty}^{\infty} h_2(t)\,dt = A_1 - A_2 \tag{14}$$

where $A_1$ and $A_2$ represent the areas under the two component functions.

We can place an upper bound on the magnitude of the transfer function if we assume that $H_2(s)$ goes to zero (dies out) before $H_1(s)$ decreases from its maximum value; that is,

$$H_{max} \leq H_1(0) = \int_{-\infty}^{\infty} h_1(t)\,dt = A_1 \tag{15}$$

We now have two rules of thumb for high-frequency enhancement filters composed of the difference of two pulses:

$$H(0) = A_1 - A_2 \quad \text{and} \quad H_{max} \leq A_1 \tag{16}$$

If $h_1(t)$ is an impulse (recall Figure 11-9), then equality holds in Eq. (16).



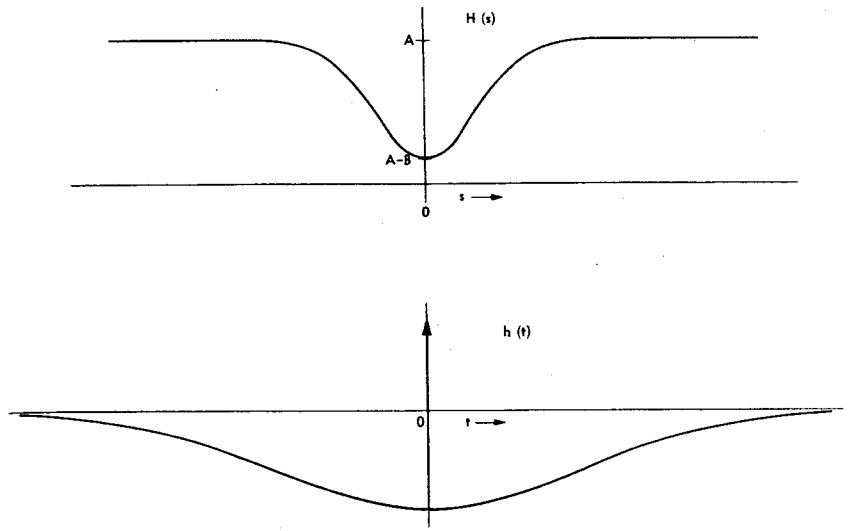



**Figure 11-9** The Gaussian highpass filter

### Low-Frequency Response

In this section, we examine the effect a filter has upon large constant areas within an image.

Assume the impulse response $g(t)$ is duration-limited, that is, zero outside a finite interval. Assume also that the input signal $f(t)$ is constant over an interval larger than the duration of $g(t)$. This situation is shown in Figure 11-11. The output of the system is given by the convolution integral

$$h(x) = \int_{-\infty}^{\infty} f(\tau)g(x - \tau)\,d\tau \tag{17}$$

Over the interval of interest, however, the input signal is constant and Eq. (17) becomes

$$h(x) = \int_{-\infty}^{\infty} cg(x - \tau)\,d\tau = c\int_{-\infty}^{\infty} g(\tau)\,d\tau \tag{18}$$

Notice that if we substitute $s = 0$ into the definition of the Fourier transform, we have

$$G(0) = \int_{-\infty}^{\infty} g(t)\,dt$$

which means

$$h(x) = cG(0) \tag{19}$$

Thus if $G(0) = 1$, the filter will not change the amplitude of large constant areas of

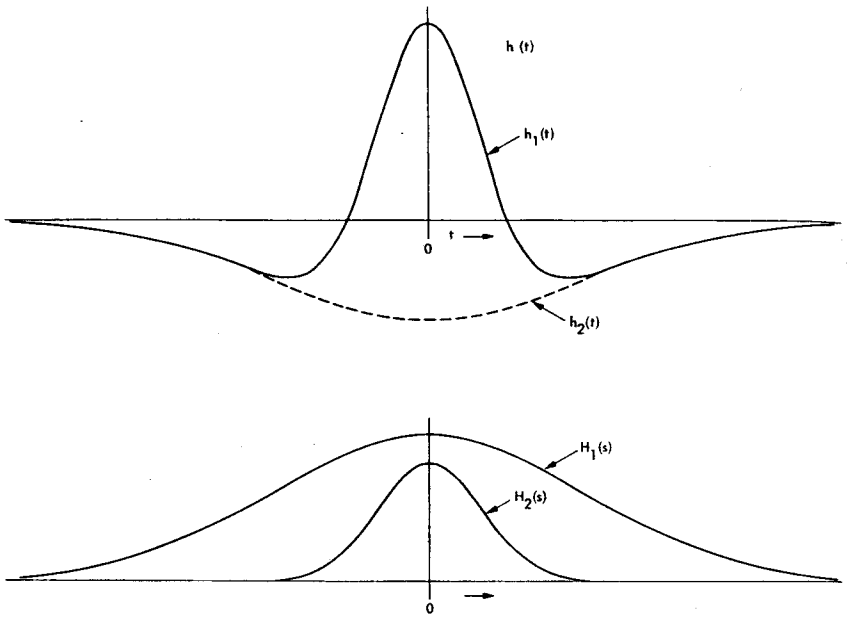



**Figure 11-10** The general highpass filter

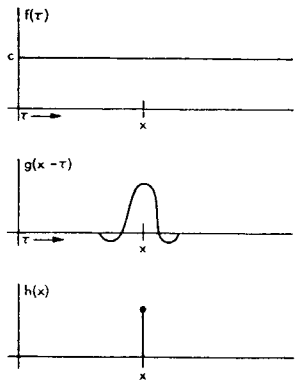



**Figure 11-11** Low-frequency response



$f(x)$. Generalizing to two dimensions, this means that the filter does not change the contrast of large flat areas within the input image. If $G(0)$ does not equal one, it becomes a gain factor controlling the overall amplitude relationship between large components of $h(t)$ and $f(t)$.

## OPTIMAL FILTER DESIGN

In this section, we develop techniques to design filters that are, in some sense, optimal for the job for which they are intended. We shall accomplish this by first establishing a criterion of "goodness" and then maximizing that criterion by proper selection of the impulse response of the filter.

The history of digital image processing has seen considerable filter design done, the way flying was done in World War I, "by the seat of the pants." Filters have been chosen for reasons of computational simplicity, past success, convenience, aesthetic appeal, and whim. Such filter design bears the unwholesome label "suboptimal," with all its negative connotations. It almost never produces the best filter for the job, and it can be dangerous.

Suboptimal filters, particularly those easy to implement by computer, can introduce artifacts into an image, frequently without warning. Filters involving the rectangular pulse in the spatial domain, a favorite of computer programmers, have a most unsavory spectrum due to the infinite undulations of the sin $(x)/x$ function. Users of such filters are often plagued by "ringing" and other artifactual phenomena in the opposite domain. They frequently regard these undesirable characteristics as indigenous to digital processing, or lament the lack of computer power necessary to do the job correctly.

In this section, we develop design techniques for optimal filters and show that they are, in general, quite well behaved. It is hoped that the reader, armed with this knowledge, can intelligently trade off optimality for computational simplicity without courting disastrous artifacts.

In this section, we review the concept of an ergodic random variable and develop design techniques for two optimal filters. They are the Wiener estimator (Refs. 1, 2, 3, 4), which is optimal for recovering an unknown signal from additive noise, and the matched detector (Refs. 4, 5, 6), which is optimal for finding a known signal buried in additive noise. Even if the reader never designs an optimal filter, these two developments will sharpen his insight considerably.

### Random Variables

In previous chapters, we have referred to the concept of a random variable, particularly for describing system noise. Since random variables play a major role in the following development, we consider them now in some detail.

We use the term *random noise* to describe an unknown contaminating signal. The word *random* is a euphemism for our incomplete knowledge. This ignorance

results from dealing with a process, the physics of which is not well understood, or with a process too complicated to analyze in detail.

When we record a signal, we know that, during the recording process, an undesired contaminating signal will appear superimposed upon (added to) the desired signal. Though we might know the origin of the noise, we cannot express its functional form mathematically. After observing the noise for a period of time, we may develop a partial knowledge of it and be able to characterize some aspects of its behavior, though we can never predict it in detail. Thus the concept of a random variable becomes a useful tool in dealing with noise.

We may think of a random variable as follows. Consider an ensemble of functions consisting of infinitely many member functions. When we make our recording, one of those member functions emerges to contaminate our record. We have no way of knowing which member function will appear, but we can make general statements about the ensemble as a group. In this way, we can express our partial knowledge of the contaminating signal.

**Ergodic Random Variables.** In the remainder of this book it is sufficient to concern ourselves only with random variables that are *ergodic*. The definition of the ergodicity property can be approached as follows. There are two ways by which we can compute averages of a random variable. We can compute a "time average" by integrating a particular member function over all time, or we can average together the values of all member functions evaluated at some particular time. The latter technique produces an "ensemble average" at some point in time. A random variable is ergodic if and only if (1) the time averages of all member functions are equal, (2) the ensemble average is constant with time, and (3) the time average and the ensemble average are numerically equal. Thus, for ergodic random variables, time averages and ensemble averages are interchangeable.

In Chapter 7 we introduced the expectation operator $\mathcal{E}\{x(t)\}$ which denotes the ensemble average of the random variable $x$ computed at time $t$. Under the ergodicity property, $\mathcal{E}\{x(t)\}$ also denotes the value obtained when $x(t)$ is averaged over time,

$$\mathcal{E}\{x(t)\} = \int_{-\infty}^{\infty} x(t)\, dt \tag{20}$$

Equation (128) of Chapter 10 defines the autocorrelation function as a time average. For an ergodic random variable, the autocorrelation function is the same for all member functions, and thus characterizes the ensemble. Therefore, when we say $n(t)$ is an ergodic random variable, we mean it is an unknown function that has a known autocorrelation function. This represents the state of our partial knowledge of $n(t)$. Since the autocorrelation function of $n(t)$

$$R_n(\tau) = \int_{-\infty}^{\infty} n(t)n(t+\tau)\, dt \tag{21}$$

is known, its power spectrum

$$P_n(s) = \mathcal{F}\{R_n(\tau)\} \tag{22}$$



is also known. This means we know the amplitude spectrum of $n(t)$ but do not know its phase spectrum. Indeed, the ensemble is composed of infinitely many functions that differ only in their phase spectra. Any real, even, nonnegative function can be the power spectrum of a random variable, and any real, even function that has a nonnegative spectrum can be the autocorrelation function of a random variable.

Fortunately, ergodic random variables model commonly encountered random signals quite well. For example, repeated observations of "white noise" sources show that the measured power spectrum is constant with frequency to a good approximation.

### The Wiener Estimator

Suppose we have an observed signal $x(t)$, which is composed of a desired signal $s(t)$ contaminated by an additive noise function $n(t)$. We would like to design a linear filter to reduce the contaminating noise as much as possible and restore the signal as closely as possible to its original form. The configuration is shown in Figure 11-12. The

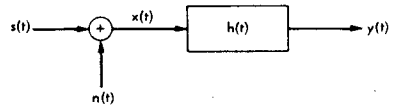



**Figure 11-12** Model for the Wiener estimator

impulse response is $h(t)$, and the output of the filter is $y(t)$. We want to select the impulse response so that the output signal $y(t)$ will be as close as possible to $s(t)$. Ideally, we would like $y(t)$ to be equal to $s(t)$, but in general a linear filter is not powerful enough to recover a contaminated signal exactly. What we shall do instead is select the impulse response $h(t)$ so that $y(t)$ will be as close as possible to $s(t)$.

Before we begin, we must decide what knowledge we have about $s(t)$ and $n(t)$. If we know nothing at all about the signal or the noise, we cannot even start on the problem. At the other extreme, if we know one or both of the signals exactly, the solution is trivial. For the purposes of the following analysis, we shall assume that both $s(t)$ and $n(t)$ are ergodic random variables and thus have known power spectra. This means that, although we do not know $n(t)$ exactly, we do know that it comes from an ensemble of functions all having the same autocorrelation function and, hence, the same power spectrum. The same restriction applies to $s(t)$. Furthermore, we assume that either we know the power spectra *a priori* or we can capture samples of $s(t)$ and $n(t)$ and determine their power spectra, which are, in turn, representative of their respective ensembles.

**Optimality Criterion.** Before we begin the development of the optimal filter, we must establish an objective criterion of optimality. Since asking for $y(t) = s(t)$ is in general asking too much of a linear filter, we shall ask instead for the best job possible under the circumstances. As a criterion of optimality, we shall use the mean-square error.

No matter what $h(t)$ is, optimal or not, we will obtain an output $y(t)$ in response to an input $s(t)$. We define the error signal at the output of the filter as

$$e(t) = s(t) - y(t) \tag{23}$$

that is, the amount by which the actual output differs from the desired output as a function of time. If the impulse response $h(t)$ is well chosen, the error signal will be, on the average, less than it would be with a poor choice of $h(t)$. As a measure of the average error, we use the mean-square error given by

$$\text{MSE} = \mathcal{E}\{e^2(t)\} = \int_{-\infty}^{\infty} e^2(t)\,dt \tag{24}$$

The latter equality holds because the error signal, a linear combination of ergodic random variables, is itself an ergodic random variable.

Notice that $e^2(t)$ is positive for both positive and negative errors. Also, squaring the error causes large errors to be penalized more severely than small errors. For these reasons, minimizing the mean-square error is an intuitively satisfactory choice of an optimality criterion. While other criteria, such as absolute error, could be used, they considerably complicate the analysis and provide, for our purposes, little or no advantage.

**The Mean-Square Error.** We now approach the problem as follows: Given the power spectra of $s(t)$ and $n(t)$, we must determine the impulse response $h(t)$ that minimizes the mean-square error. Notice that the mean-square error is a functional of $h(t)$, the impulse response, since a function $h(t)$ maps into a real number MSE. The branch of mathematics concerned with functional minimization is the calculus of variations, which we shall use. We shall obtain a functional expression for MSE in terms of $h(t)$, then find an expression for the optimal (minimizing) $h(t)$ in terms of known power spectra, and finally develop an expression for the MSE that results when the optimal $h(t)$ is used. This latter step will show how well the optimal filter works.

We begin by expanding the mean-square error in Eq. (24).

$$\text{MSE} = \mathcal{E}\{e^2(t)\} = \mathcal{E}\{[s(t) - y(t)]^2\} = \mathcal{E}\{s^2(t) - 2s(t)y(t) + y^2(t)\} \tag{25}$$

Since the expectation is an integral operator [Eq. (20)], we can write

$$\text{MSE} = \mathcal{E}\{s^2(t)\} - 2\mathcal{E}\{s(t)y(t)\} + \mathcal{E}\{y^2(t)\} = T_1 + T_2 + T_3 \tag{26}$$

where $T_1$, $T_2$, and $T_3$ are introduced so that we may consider the three terms separately. Writing $T_1$ as an integral,

$$T_1 = \mathcal{E}\{s^2(t)\} = \int_{-\infty}^{\infty} s^2(t)\,dt = R_s(0) \tag{27}$$

we recognize it as the $\tau = 0$ point on the (known) autocorrelation function of $s(t)$.

Writing $y(t)$ as the convolution of $x(t)$ and $h(t)$ allows us to expand the second term as

$$T_2 = -2\mathcal{E}\{s(t)\int_{-\infty}^{\infty} h(\tau)x(t-\tau)\,d\tau\} \tag{28}$$

Since the expectation operator is actually an integral over time, we can rearrange Eq. (28) to produce

$$T_2 = -2\int_{-\infty}^{\infty} h(\tau)\mathcal{E}\{s(t)x(t-\tau)\}\,d\tau \tag{29}$$

Now we recognize the expectation inside the integral as the cross-correlation function of $s(t)$ and $x(t)$ and write

$$T_2 = -2\int_{-\infty}^{\infty} h(\tau)R_{sx}(\tau)\,d\tau \tag{30}$$

We can expand $T_3$ as the expectation of the product of two convolutions

$$T_3 = \mathcal{E}\left\{\int_{-\infty}^{\infty} h(\tau)x(t-\tau)\,d\tau \int_{-\infty}^{\infty} h(u)x(t-u)\,du\right\} \tag{31}$$

which may be rearranged as before to yield

$$T_3 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau)h(u)\mathcal{E}\{x(t-\tau)x(t-u)\}\,d\tau\,du \tag{32}$$

If, inside the expectation operator, we make the variable substitution $v = t - u$, that factor becomes

$$\mathcal{E}\{x(t-\tau)x(t-u)\} = \mathcal{E}\{x(v+u-\tau)x(v)\} \tag{33}$$

which is simply the autocorrelation function of $x(t)$ evaluated at the point $u - \tau$. Now the third term can be written as

$$T_3 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau)h(u)R_x(u-\tau)\,d\tau\,du \tag{34}$$

The mean-square error of Eq. (26) can now be written as

$$\text{MSE} = R_s(0) - 2\int_{-\infty}^{\infty} h(\tau)R_{sx}(\tau)\,d\tau + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau)h(u)R_x(u-\tau)\,d\tau\,du \tag{35}$$

This is the mean-square error in terms of the filter's impulse response and known autocorrelation and cross-correlation functions of the two input signal components. As expected, MSE is a functional of $h(t)$. We now wish to select the particular function $h(t)$ that causes MSE to take on its minimum value.

**Minimizing MSE.** Let us denote by $h_o(t)$ the particular function that minimizes MSE. In general, an arbitrary $h(t)$ will differ from the optimal $h_o(t)$, and we can define a function $g(t)$ to account for this variation from the optimal; that is,

$$h(t) = h_o(t) + g(t) \tag{36}$$

where $h(t)$ is an arbitrarily chosen (suboptimal) impulse response function and $g(t)$ is chosen to make the equality hold. The reason for this seemingly unnecessary complication is not obvious now, but it will allow us to establish a necessary condition upon $h_o(t)$.

If we substitute the definition for $g(t)$ in Eq. (36) into the MSE equation [Eq. (35)], we obtain

$$\begin{aligned}\text{MSE} = R_s(0) &- 2\int_{-\infty}^{\infty} [h_o(\tau) + g(\tau)]R_{sx}(\tau)\,d\tau \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [h_o(\tau) + g(\tau)]\,[h_o(u) + g(u)]R_x(u-\tau)\,d\tau\,du\end{aligned} \tag{37}$$

This expression can be expanded, producing seven terms

$$\begin{aligned}\text{MSE} = R_s(0) &- 2\int_{-\infty}^{\infty} h_o(\tau)R_{xs}(\tau)\,d\tau + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_o(\tau)h_o(u)R_x(u-\tau)\,d\tau\,du \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_o(\tau)g(u)R_x(u-\tau)\,d\tau\,du + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_o(u)g(\tau)R_x(u-\tau)\,d\tau\,du \\ &- 2\int_{-\infty}^{\infty} g(\tau)R_{xs}(\tau)\,d\tau + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\tau)g(u)R_x(u-\tau)\,d\tau\,du \end{aligned} \tag{38}$$

Comparing the first three terms with Eq. (35) we see that their sum represents the mean-square error that results when the optimal impulse response $h_o(t)$ is used. We denote this value by $\text{MSE}_o$. Since the autocorrelation function $R_x(u-\tau)$ is an even function, the fourth and fifth terms of Eq. (38) are equal. We can combine them with the sixth term and write Eq. (38) as

$$\begin{aligned}\text{MSE} = \text{MSE}_o &+ 2\int_{-\infty}^{\infty} g(u)\left[\int_{-\infty}^{\infty} h_o(\tau)R_x(u-\tau)\,d\tau - R_{xs}(u)\right]du \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u)g(\tau)R_x(u-\tau)\,du\,d\tau = \text{MSE}_o + T_4 + T_5\end{aligned} \tag{39}$$

where $T_4$ and $T_5$ are introduced for compactness of notation.

We shall now show that the term $T_5$ is nonnegative. Writing the autocorrelation function $R_x(u-\tau)$ as an integral produces

$$T_5 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u)g(\tau)\int_{-\infty}^{\infty} x(t-\tau)x(t-u)\,dt\,du\,d\tau \tag{40}$$

which may be rearranged to yield

$$T_5 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u)x(t-u)\,du\int_{-\infty}^{\infty} g(\tau)x(t-\tau)\,d\tau\,dt \tag{41}$$

If we define $z(t)$ as the function that results from convolving $g(t)$ with $x(t)$, we can recognize Eq. (41) as

$$T_5 = \int_{-\infty}^{\infty} z^2(t)\,dt \geq 0 \tag{42}$$

which can never be negative.

Returning to the MSE, we can write Eq. (39) as

$$\text{MSE} = \text{MSE}_o + 2\int_{-\infty}^{\infty} g(u)\left[\int_{-\infty}^{\infty} h_o(\tau)R_x(u-\tau)\,d\tau - R_{xs}(u)\right]du + T_5 \tag{43}$$

where $\text{MSE}_o$ is the mean-square error under optimal conditions and $T_5$ cannot be negative. We wish to establish a condition on $h_o(\tau)$ that will ensure that $\text{MSE}_o$ is the smallest value that MSE can possibly have. One way to do this is to make the quantity in brackets be zero for all values of $u$. This makes $T_4$ drop out of Eq. (43) and guarantees that $\text{MSE}_o \leq \text{MSE}$. However, we still must make sure that condition is both necessary and sufficient to optimize the filter.

Suppose that the term in brackets in Eq. (43) were nonzero for some values of $u$. Since $g(u)$ is an abritrary function, it could take on large negative values where the bracketed term was positive and vice versa. The integral in $T_4$ would then take on a large negative value and MSE would become smaller than $\text{MSE}_o$. Since this violates our definition, we conclude that it is a necessary condition that the bracketed term in Eq. (43) must be identically zero. This means that



$$R_{xs}(\tau) = \int_{-\infty}^{\infty} h_o(u)R_x(u-\tau)\,du \tag{44}$$

is a necessary condition in order for the mean-square error to be minimized. Now the complication introduced in Eq. (36) has paid off by giving us a necessary condition for the optimal filter.

It is easy to see that the condition of Eq. (44) is also sufficient to optimize the filter, that is, that no additional conditions are required. Since the necessary condition causes $T_4$ to drop out of Eq. (43), it becomes

$$\text{MSE} = \text{MSE}_o + T_5 \qquad T_5 \geq 0 \tag{45}$$

from which it is clear that

$$\text{MSE} \geq \text{MSE}_o \tag{46}$$

Thus Eq. (44) defines the impulse response of the linear estimator that is optimal in the mean-square sense.

Notice that the right-hand side of Eq. (44) is a convolution integral, which can be written as

$$R_{xs}(\tau) = h_o(u) * R_x(u) \tag{47}$$

It relates the optimal impulse response to the autocorrelation of the input signal and the cross-correlation of the input and the desired signal.

It is easy to show that, for any linear system, the cross-correlation between input and output is given by

$$R_{xy}(\tau) = h(u) * R_x(u) \tag{48}$$

where $R_x(u)$ is the autocorrelation function of the input signal [see Chapter 13, Eq. (57)]. Comparing this with Eq. (47) illustrates that the Wiener filter makes the input/output cross-correlation function equal to the signal/signal-plus-noise cross-correlation function.

If we take the Fourier transform of both sides of Eq. (47), we are left with

$$P_{xs}(s) = H_o(s)P_x(s) \tag{49}$$

which implies that

$$H_0(s) = \frac{P_{xs}(s)}{P_x(s)} \tag{50}$$

is the frequency domain specification of the Wiener estimator.

**Wiener Filter Design.** Equation (50) implies that we can design a Wiener estimator in the following way: (1) First, digitize a sample of the input signal $s(t)$. (2) Autocorrelate the input sample to produce an estimate of $R_x(\tau)$. (3) Compute the Fourier transform of $R_x(\tau)$ to produce $P_x(s)$. (4) Obtain and digitize a sample of the signal in the absence of noise. (5) Cross-correlate the signal sample with the input sample to estimate $R_{xs}(\tau)$. (6) Compute the Fourier transform of $R_{xs}(\tau)$ to produce $P_{xs}(s)$. (7)

Compute the transfer function of the optimal filter by Eq. (50). (8) If the filter is to be implemented by convolution, compute the inverse Fourier transform of $H_o(s)$ to produce the impulse response $h_o(t)$ of the optimum linear estimator.

If it is impossible or impractical to obtain samples of the noise-free signal and the input signal, one could assume a functional form for the correlation functions or the power spectra required in Eq. (50). For example, white noise has a constant power spectrum, and some functional form might be assumed for the power spectrum of the desired signal.

**Uncorrelated Signal and Noise.** The autocorrelation functions in Eq. (45) and the power spectra in Eq. (50) are somewhat difficult to visualize and interpret. The situation is improved considerably, however, if we assume that the desired signal and the noise are uncorrelated. By definition, this means that

$$\mathcal{E}\{s(t)n(t)\} = \mathcal{E}\{s(t)\}\mathcal{E}\{n(t)\} \tag{51}$$

We can transform the numerator of $H_o(s)$ [Eq. (50)] and write

$$R_{xs}(\tau) = \mathcal{E}\{x(t)s(t+\tau)\} = \mathcal{E}\{[s(t)+n(t)]s(t+\tau)\} \tag{52}$$

or

$$R_{xs}(\tau) = \mathcal{E}\{s(t)s(t+\tau)\} + \mathcal{E}\{n(t)s(t+\tau)\} \tag{53}$$

In view of Eq. (51) we can write

$$R_{xs}(\tau) = R_s(\tau) + \mathcal{E}\{n(t)\}\mathcal{E}\{s(t+\tau)\} = R_s(\tau) + \int_{-\infty}^{\infty} n(t)\,dt \int_{-\infty}^{\infty} s(t+\tau)\,dt \tag{54}$$

or

$$R_{xs}(\tau) = R_s(\tau) + N(0)S(0) \tag{55}$$

A similar exercise in the denominator of Eq. (50) produces

$$R_x(\tau) = R_s(\tau) + R_n(\tau) + 2S(0)N(0) \tag{56}$$

Then Eq. (50) becomes

$$H_o(s) = \frac{P_s(s) + N(0)S(0)\,\delta(s)}{P_s(s) + P_n(s) + 2N(0)S(0)\,\delta(s)} \tag{57}$$

or, ignoring zero frequency,

$$H_o(s) = \frac{P_s(s)}{P_s(s) + P_n(s)} \qquad s \neq 0 \tag{58}$$

Notice that if either the signal or noise has zero mean value, then Eq. (58) is valid for all frequencies, including zero. If both signal and noise have nonzero mean values, then

$$H_o(0) = \tfrac{1}{2} \tag{59}$$

Recall that Eq. (35) gives the mean-square error at the filter output. If we install the optimality condition of Eq. (44) in the third term of Eq. (35), it combines with the second term, leaving

$$\text{MSE}_o = R_s(0) - \int_{-\infty}^{\infty} h_o(\tau)R_{xs}(\tau)\,d\tau \tag{60}$$



a simple expression for the mean-square error of the optimal filter. With uncorrelated zero mean noise, Eq. (55) suggests we can replace $R_{xs}(\tau)$ with $R_s(\tau)$, which is the inverse Fourier transform of $P_s(s)$. Then Eq. (60) becomes

$$\text{MSE}_o = R_s(0) - \int_{-\infty}^{\infty} h_o(\tau)\mathfrak{F}^{-1}\{P_s(s)\}\,d\tau \tag{61}$$

Writing out the inverse transformation and rearranging integrals produces

$$\text{MSE}_o = R_s(0) - \int_{-\infty}^{\infty} P_s(s) \int_{-\infty}^{\infty} h_o(\tau)e^{j2\pi s\tau}\,d\tau\,ds \tag{62}$$

Recognizing the first term and the second integral as Fourier transforms allows us to write

$$\text{MSE}_o = \int_{-\infty}^{\infty} P_s(s)\,ds - \int_{-\infty}^{\infty} P_s(s)H_o(-s)\,ds \tag{63}$$

Since the transfer function $H_o(s)$ is even, the minus sign on its argument can be ignored. We now substitute Eq. (58) and obtain

$$\text{MSE}_o = \int_{-\infty}^{\infty} P_s(s)\,ds - \int_{-\infty}^{\infty} P_s(s)\frac{P_s(s)}{P_s(s)+P_n(s)}\,ds \tag{64}$$

which may be rearranged to yield

$$\text{MSE}_o = \int_{-\infty}^{\infty} \frac{P_s(s)P_n(s)}{P_s(s)+P_n(s)}\,ds \tag{65}$$

the frequency domain expression for mean-square error in the uncorrelated case.

Figure 11-13 illustrates the frequency domain behavior of the Wiener filter in the uncorrelated case. Notice that the magnitude of the transfer function $H_o(s)$ is bounded by 0 and 1. Also, the transfer function decreases with a decrease in signal power spectrum or an increase in noise power spectrum. When the signal-to-noise ratio is high, the transfer function approaches unity, passing all the energy in the signal. When the signal-to-noise ratio is low, however, the transfer function declines toward zero.

When we assumed that our knowledge of $s(t)$ and $n(t)$ was limited to power spectra, we admitted we had no phase information. Notice that the transfer function $H_o(s)$ is real and even and thus has no phase shift.

The actual mean-square error at the output, which indicates how successfully the filter is able to recover the signal from the contaminating noise, is given by Eq. (65). The integrand is plotted in Figure 11-13. Notice that the contributions to MSE occur in frequency bands where both the signal and noise power spectra are nonzero.

Figure 11-14 illustrates the case where signal and noise are separable in the frequency domain. In this case, the Wiener estimator passes the signal in its entirety and discriminates completely against the noise.

The case of a band-limited signal imbedded in white noise is illustrated in Figure 11-15. If the signal power spectrum is constant, the mean-square error is proportional to the signal bandwidth, $s_2 - s_1$.

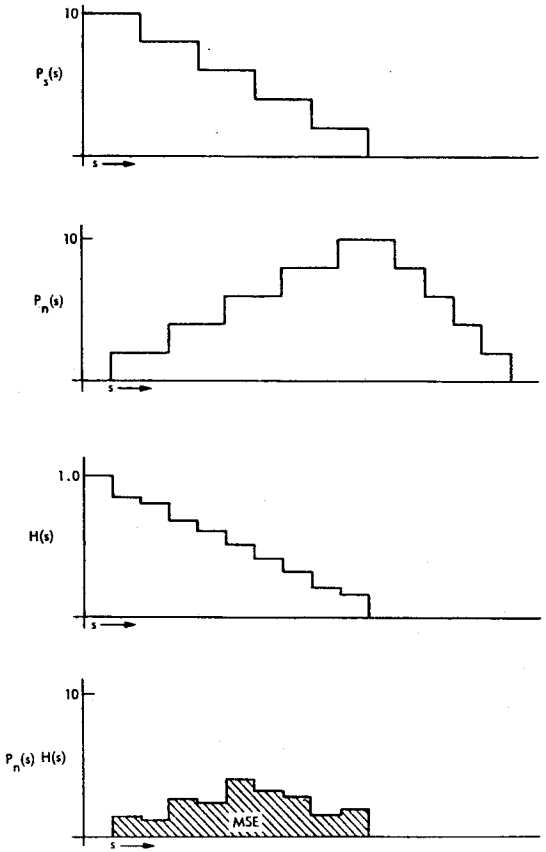



**Figure 11-13** The Wiener filter transfer function

If the signal-to-noise ratio is low, Eq. (65) reduces to approximately

$$\text{MSE}_o \approx \int_{-\infty}^{\infty} P_s(s)\,ds = \int_{-\infty}^{\infty} |S(s)|^2\,ds \tag{66}$$

which is, by Rayleigh's theorem,

$$\text{MSE}_o \approx \int_{-\infty}^{\infty} s^2(t)\,dt = R_s(0) = \text{energy} \tag{67}$$

Thus, in this case, the mean-square error is proportional to the energy in the signal.



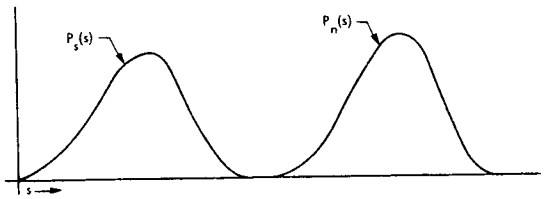



**Figure 11-14** Separable signal and noise

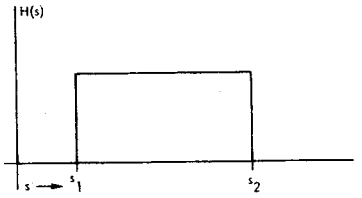



**Figure 11-15** Band-limited signal

**Wiener Deconvolution.** Ordinary deconvolution, as previously discussed, does not account for noise. Thus deconvolution transfer functions, which often take on extremely large magnitudes at high frequencies, are not practical when noise is present. Figure 11-16 illustrates the situation when deconvolution is followed by a Wiener filter. The

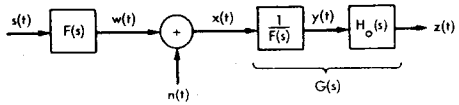



**Figure 11-16** Wiener deconvolution

desired signal $s(t)$ is first degraded by a linear system with impulse response $f(t)$. The output of the filter is corrupted by a noise source $n(t)$ to form the observed signal $x(t)$. It is desired to design a linear filter $g(t)$ that will simultaneously deconvolve the undesired impulse response $f(t)$ and discriminate against the noise. In Figure 11-16, $g(t)$ is illustrated as a concatenation of a deconvolution filter and a Wiener filter with impulse response $h_o(t)$.

Since the deconvolution filter is known, it only remains to determine the impulse response $h_o(t)$ before combining the two linear filters to produce $g(t)$.

The configuration in Figure 11-16 implies that the spectrum of the observed signal is given by

$$X(s) = F(s)S(s) + N(s) \tag{68}$$

Furthermore, assuming $F(s)$ has no zeros, the input spectrum to the Wiener filter is

$$Y(s) = S(s) + \frac{N(s)}{F(s)} = S(s) + K(s) \tag{69}$$

Equation (58) implies, for uncorrelated signal and noise sources, that the Wiener filter transfer function is given by

$$H_o(s) = \frac{P_s(s)}{P_s(s) + P_k(s)} = \frac{|S(s)|^2}{|S(s)|^2 + \left|\frac{N(s)}{F(s)}\right|^2} \tag{70}$$

Thus the transfer function $G(s)$ of the optimal deconvolution filter in the mean-square sense is given by

$$G(s) = \frac{H_o(s)}{F(s)} = \frac{1}{F(s)}\left[\frac{P_s(s)}{P_s(s) + P_k(s)}\right] = \frac{F^*(s)P_s(s)}{|F(s)|^2 P_s(s) + P_n(s)} \tag{71}$$

### The Matched Detector

We now consider a filter that is optimal for a different purpose. Whereas the Wiener filter is optimal for recovering an unknown signal from noise, the matched detector is optimal for locating a known signal in a noisy background (Refs. 4, 5, 6). The matched filter is designed to "detect" the occurrence of a signal of prescribed form in a noisy background. This contrasts with the Wiener filter, which is designed to "estimate" what the signal was before it was contaminated with noise.

The model for the development of the matched detector is shown in Figure 11-17. A signal $m(t)$ is contaminated by additive noise $n(t)$ to form the observed signal

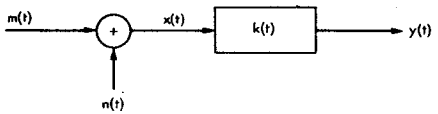



**Figure 11-17** Model for the matched detector

$x(t)$, which is input to the linear filter having impulse response $k(t)$, producing the output $y(t)$. We wish to use the filter to detect the presence or absence of $m(t)$. That is to say, we shall monitor $y(t)$ to detect the occurrence of $m(t)$, a specified signal of known form. We wish to select the impulse response $k(t)$ to make this job easy.

For the system in Figure 11-17,

$$y(t) = [m(t) + n(t)] * k(t) = m(t) * k(t) + n(t) * k(t) \tag{72}$$

which means that the system in Figure 11-18 is equivalent. It makes no difference whether $m(t)$ and $n(t)$ are summed before or after passing through the filter. We define the component outputs as

$$u(t) = m(t) * k(t) \quad \text{and} \quad v(t) = n(t) * k(t) \tag{73}$$

Now $u(t)$ is the filtered signal and $v(t)$ is the filtered noise.



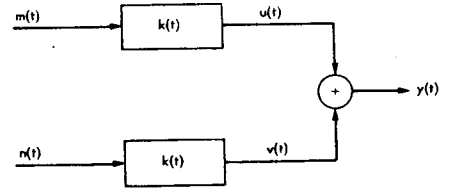



**Figure 11-18** An equivalent model

As with the Wiener filter, we must first stipulate what knowledge we have about the signal and noise and establish a criterion of optimality. Suppose we know the functional form of $m(t)$, but we do not know at what point in time it occurs. The classical application of the matched detector has been the detection of reflected radar pulses. In this case, the reflected pulse, identical to the transmitted pulse, is known in form but not in time of arrival. In digital image processing, the matched detector is useful for locating known features in a noisy image.

As with the Wiener filter, we shall assume that the noise is an ergodic random variable of known power spectrum. We wish to design $k(t)$ so that, by observing the output, we may best be able to detect the signal when it does occur.

**Optimality Criterion.** As a measure of performance for the filter, we shall use the average signal-to-noise power ratio at the output, evaluated at time $t_o$,

$$\rho = \frac{\mathcal{E}\{u^2(t_0)\}}{\mathcal{E}\{v^2(t_0)\}} \tag{74}$$

assuming that the signal $m(t)$ occurs at $t = 0$.

Since the system is shift invariant, if the signal should choose to occur at some time $t_1$, then the signal to noise power ratio at the output will be maximized at $t_0 + t_1$. Thus $t_0$ allows us to introduce offset, if desired, between the occurrences of the signal and the output pulse. Ordinarily, for filter design purposes, the signal $m(t)$ is some relatively narrow function located at the origin, and $t_0$ is set to zero. Then, when $m(t - t_1)$ arrives at time $t_1$, the amplitude of the filter output becomes large. Before and after, in the absence of signal, the output amplitude is relatively small.

Clearly, if $\rho$ is large, the amplitude of the output $y(t)$ will be highly dependent on the presence or absence of $m(t)$ and relatively insensitive to fluctuations in the noise $n(t)$. Thus, as a criterion for optimality of $k(t)$, we shall choose the maximization of $\rho$.

It is important to note that this criterion of optimality makes no guarantee that the output $y(t)$ will resemble $m(t)$. However, since we know the functional form of $m(t)$, we are not interested in fidelity of reproduction, as we were in the case of the Wiener filter. Instead, we want the output to be large when $m(t)$ is present and small when it is not.

Since $u(t)$ is deterministic, we can drop the expectation operator in the numerator and write Eq. (74) as

$$\rho = \frac{u^2(t_0)}{\mathcal{E}\{v^2(t)\}} = \frac{[m(t) * k(t)]^2}{\mathcal{E}\{[n(t) * k(t)]^2\}} = \frac{[\mathcal{F}^{-1}\{M(s)K(s)\}]^2}{\mathcal{E}\{[n(t) * k(t)]^2\}} = \frac{\rho_n}{\rho_d} \tag{75}$$

We begin by expanding the denominator as a product of two convolution integrals:

$$\rho_d = \mathcal{E}\left\{\int_{-\infty}^{\infty} k(q)n(t-q)\,dq \int_{-\infty}^{\infty} k(\tau)n(t-\tau)\,d\tau\right\} \tag{76}$$

Since the expectation is an integral over time and the impulse response $k(t)$ is not a random signal, we can rearrange the integrals in Eq. (76) to produce

$$\rho_d = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k(q)k(\tau)\mathcal{E}\{n(t-q)n(t-\tau)\}\,dq\,d\tau \tag{77}$$

We recognize the expectation factor within the integral as the autocorrelation function $R_n(\tau - q)$ of the noise, which is, in turn, the inverse Fourier transform of $P_n(s)$, the noise power spectrum. Thus

$$\mathcal{E}\{n(t-q)n(t-\tau)\} = R_n(\tau - q) = \int_{-\infty}^{\infty} P_n(s)e^{j2\pi s(\tau-q)}\,ds \tag{78}$$

which makes the denominator of $\rho$

$$\rho_d = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k(q)k(\tau)\int_{-\infty}^{\infty} P_n(s)e^{j2\pi s(\tau-q)}\,ds\,dq\,d\tau \tag{79}$$

Now we can factor the exponential and rearrange the integrals to produce

$$\rho_d = \int_{-\infty}^{\infty} P_n(s)\left[\int_{-\infty}^{\infty} k(q)e^{-j2\pi sq}\,dq \int_{-\infty}^{\infty} k(\tau)e^{j2\pi s\tau}\,d\tau\right]ds \tag{80}$$

The term in brackets is the product of two inverse Fourier transforms, namely $K(s)K(-s)$. Furthermore, since the impulse response $k(t)$ is a real function, the transfer function $K(s)$ is Hermite and $K(-s) = K^*(s)$. Thus the term in brackets reduces to

$$K(s)K(-s) = K(s)K^*(s) = |K(s)|^2 \tag{81}$$

Substituting this into Eq. (75) and writing out the Fourier transform in the numerator allows us to write the signal-to-noise power ratio as

$$\rho = \frac{\left[\int_{-\infty}^{\infty} K(s)M(s)e^{j2\pi st_0}\,ds\right]^2}{\int_{-\infty}^{\infty} |K(s)|^2 P_n(s)\,ds} \tag{82}$$

It is this expression we wish to maximize. As with the Wiener filter, we must select a function to optimize a quantity.

**Schwartz's Inequality.** In this case, we shall make use of Schwartz's inequality. This is the mathematical result which states that

$$\int f^2(t)\,dt \int g^2(t)\,dt \geq \left[\int f(t)g(t)\,dt\right]^2 \tag{83}$$

where $f(t)$ and $g(t)$ are arbitrary real functions and the integration is performed between arbitrary limits. Our approach is to define the functions $f(t)$ and $g(t)$ in terms of factors appearing in Eq. (82) and obtain an inequality involving $\rho$. We shall then assume a form for the transfer function and show that it maximizes $\rho$. First, however, we shall prove the Schwartz inequality.



We first define a function of the variable $\lambda$ by writing

$$Q(\lambda) = \int [\lambda f(t) + g(t)]^2\,dt \geq 0 \tag{84}$$

Expanding the integrand and collecting terms produces

$$\int [\lambda f(t) + g(t)]^2\,dt = \lambda^2 \int f^2(t)\,dt + 2\lambda \int f(t)g(t)\,dt + \int g^2(t)\,dt \geq 0 \tag{85}$$

Equation (85) is a quadratic equation in the variable $\lambda$. Therefore,

$$\left[2\int f(t)g(t)\,dt\right]^2 - 4\int f^2(t)\,dt \int g^2(t)\,dt \leq 0 \tag{86}$$

or

$$\left[\int f(t)g(t)\,dt\right]^2 \leq \int f^2(t)\,dt \int g^2(t)\,dt \tag{87}$$

thus proving Eq. (83).

**A Necessary Condition.** We shall now use Schwartz's inequality to obtain a condition upon the signal-to-noise ratio $\rho$. First we define two functions

$$f(s) = e^{j2\pi st_0}K(s)\sqrt{P_n(s)} \tag{88}$$

and

$$g(s) = \frac{M(s)}{\sqrt{P_n(s)}} \tag{89}$$

Their product is

$$f(s)g(s) = e^{j2\pi st_0}K(s)M(s) \tag{90}$$

and their squared magnitudes are

$$|f(s)|^2 = |K(s)|^2 P_n(s) \tag{91}$$

and

$$|g(s)|^2 = \frac{|M(s)|^2}{P_n(s)} \tag{92}$$

If we substitute the functions defined in Eqs. (88) and (89) into Schwartz's inequality, using $s$ as the variable of integration, we obtain

$$\left|\int_{-\infty}^{\infty} e^{j2\pi st_0}K(s)M(s)\,ds\right|^2 \leq \left[\int_{-\infty}^{\infty} |K(s)|^2 P_n(s)\,ds\right]\left[\int_{-\infty}^{\infty} \frac{|M(s)|^2}{P_n(s)}\,ds\right] \tag{93}$$

If we divide both sides by

$$\int_{-\infty}^{\infty} |K(s)|^2 P_n(s)\,ds \tag{94}$$

we are left with

$$\frac{\left|\int_{-\infty}^{\infty} e^{j2\pi st_0}K(s)M(s)\,ds\right|^2}{\int_{-\infty}^{\infty} |K(s)|^2 P_n(s)\,ds} \leq \frac{\int_{-\infty}^{\infty} |K(s)|^2 P_n(s)\,ds \int_{-\infty}^{\infty} \frac{|M(s)|^2}{P_n(s)}\,ds}{\int_{-\infty}^{\infty} |K(s)|^2 P_n(s)\,ds} \tag{95}$$

Recalling Eq. (82), we recognize the left side of the inequality as $\rho$. Furthermore, the

denominator on the right-hand side cancels the first term of the numerator, leaving us with

$$\rho \leq \int_{-\infty}^{\infty} \frac{|M(s)|^2}{P_n(s)}\, ds \tag{96}$$

a relatively simple upper bound on $\rho$.

Schwartz's inequality has led us to Eq. (96), which states that $\rho$ is less than or equal to an expression involving the power spectrum of the signal and that of the noise. Clearly, $\rho$ will be maximized when equality holds in Eq. (96). Since we want $\rho$ to be as large as possible, we take

$$\rho_{\max} = \int_{-\infty}^{\infty} \frac{|M(s)|^2}{P_n(s)}\, ds \tag{97}$$

as a necessary condition to maximize $\rho$.

**The Transfer Function.** We shall now assume a particular form for $K(s)$ and show that it does indeed maximize $\rho$. We assume that the optimal transfer function is given by

$$K_o(s) = Ce^{-j2\pi st_0}\frac{M^*(s)}{P_n(s)} \tag{98}$$

Substituting that assumed form into the general expression for $\rho$ [Eq. (82)] produces

$$\rho = \frac{\left|\int_{-\infty}^{\infty} Ce^{-j2\pi st_0}e^{j2\pi st_0}\frac{M^*(s)}{P_n(s)}M(s)\, ds\right|^2}{\int_{-\infty}^{\infty} C^2\frac{M^*(s)M(s)}{P_n^*(s)P_n(s)}P_n(s)\, ds} \tag{99}$$

Canceling the constants, the exponentials, and the $P_n(s)$ in the denominator reduces the expression to

$$\rho = \frac{\left|\int_{-\infty}^{\infty} \frac{M^*(s)}{P_n(s)}M(s)\, ds\right|^2}{\int_{-\infty}^{\infty} \frac{M^*(s)M(s)}{P_n^*(s)}\, ds} \tag{100}$$

Since $P_n(s)$ is real and even, $P_n^*(s) = P_n(s)$, and the numerator is the square of the denominator. Now $\rho$ reduces to

$$\rho = \int_{-\infty}^{\infty} \frac{|M(s)|^2}{P_n(s)}\, ds = \rho_{\max} \tag{101}$$

which satisfies the necessary condition for optimality of Eq. (97). This means that the transfer function assumed in Eq. (98) does indeed maximize the signal-to-noise power ratio at the output of the filter at time $t_0$ when the signal occurs at $t = 0$.

Notice that the magnitude of the transfer function

$$|K_o(s)| = |C|\frac{|M(s)|}{P_n(s)} \tag{102}$$

is proportional to the signal amplitude to noise power ratio as a function of frequency.



The arbitrary constant $C$ is not surprising, since we originally endeavored to maximize a ratio at the output.

#### Examples of the Matched Detector

In order to develop insight into the operation of the matched detector, we consider some illustrative examples under particular conditions.

**White Noise.** In the first case, let us assume that the noise $n(t)$ is spectrally white; that is,

$$P_n(s) = N_0^2 \tag{103}$$

Since $C$ in Eq. (98) is an arbitrary constant, we may set it equal to $N_0^2$, in which case the matched detector becomes

$$K_o(s) = M^*(s)e^{-j2\pi st_0} \tag{104}$$

In the time domain, the impulse response is

$$k_o(t) = \mathfrak{F}^{-1}\{K(s)\} = \int_{-\infty}^{\infty} M^*(s)e^{-j2\pi st_0}e^{j2\pi st}\, ds \tag{105}$$

Since $m(t)$ is real, $M(s)$ is Hermite and

$$k_o(t) = \int_{-\infty}^{\infty} M(-s)e^{j2\pi(-s)(t_0-t)}\, ds = m(t_0 - t) \tag{106}$$

Thus the impulse response for the white noise case is merely a reflected and shifted version of the signal itself. This filter is said to be "matched" to the signal (Ref. 5), and this name has become attached to the more general detector of Eq. (98).

The signal component of the output is given by

$$u(t) = m(t) * k_o(t) = \int_{-\infty}^{\infty} m(\tau)m(t_0 - t + \tau)\, d\tau = R_m(t_0 - t) \tag{107}$$

and the noise component by

$$v(t) = n(t) * k_o(t) = \int_{-\infty}^{\infty} n(\tau)m(t_0 - t + \tau)\, d\tau = R_{mn}(t_0 - t) \tag{108}$$

Since $k_o(t)$ in Eq. (106) is just the reflected signal we are trying to detect, the matched filter $k_o(t)$ is merely a cross-correlator. It cross-correlates the incoming signal plus noise with the known form of the desired signal. The output is

$$y(t) = u(t) + v(t) = R_m(t_0 - t) + R_{mn}(t_0 - t) \tag{109}$$

which has an autocorrelation component only when the signal is present and always a cross-correlation component. If the correlation between the signal and noise is small, then $R_{mn}(\tau)$ is small for all values of $\tau$ and the noise component at the output is small. Furthermore, the autocorrelation function $R_m(\tau)$ has a peak at $\tau = 0$ so, clearly,

$$\rho = \frac{u^2(t_0)}{\mathcal{E}\{v^2(t)\}} \tag{110}$$

is large at $t = t_0$ as desired.

**The Rectangular Pulse Detector.** As a particular example, suppose $m(t) = \Pi(t)$; that is, the matched filter is designed to detect a rectangular pulse in white noise. Suppose also that the input is $x(t) = s(t) + n(t)$, where $s(t) = \Pi(t - T)$ and $n(t)$ is white noise. Recall that the autocorrelation function of the rectangular pulse is given by

$$R_x(\tau) = \Pi(t) * \Pi(t) = \Lambda(\tau) \tag{111}$$

Now the output of the filter is

$$y(t) = R_{xm}(t) = R_{sm}(t) + R_{mn}(t) = \Lambda(t - T) + R_{mn}(t) \tag{112}$$

So, for the system shown in Figure 11-19, the components of the input and output are presented in Figure 11-20.

From Figure 11-20, we see how the matched filter discriminates against the noise while responding to the signal. The output has a peak at $t = T$, the time at which the input pulse occurs, but takes on relatively small amplitude otherwise. Thus a simple examination of the output signal indicates when the input pulse occurs.

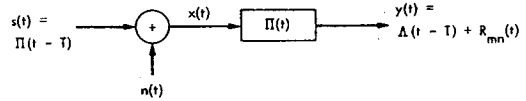



**Figure 11-19** Rectangular pulse detector

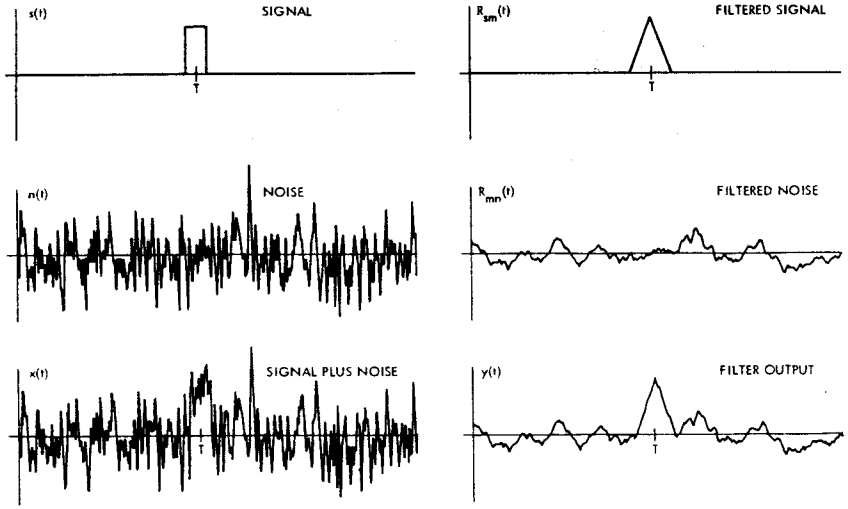



**Figure 11-20** Input and output component signals



Notice that the form or shape of the signal is not preserved by the matched detector as it was with the Wiener estimator. This is because we designed the filter to detect the existence of a particular known input signal rather than to estimate its noise-free shape.

### Comparison of the Wiener Estimator and the Matched Detector

The Wiener estimator and the matched detector are both optimal filters designed to do specific jobs. It is instructive to compare the two filters. Recall from Eq. (58) that for uncorrelated signal and noise the Wiener estimator transfer function is given by

$$H_o(s) = \frac{P_s(s)}{P_s(s) + P_n(s)} \tag{113}$$

and the mean-square error one can expect when using this filter is, from Eq. (65),

$$MSE_o = \int_{-\infty}^{\infty} \frac{P_s(s)P_n(s)}{P_s(s) + P_n(s)}\, ds \tag{114}$$

If we let $c = 1$, and $t_0 = 0$, the matched detector transfer function is

$$K_o(s) = \frac{S^*(s)}{P_n(s)} \tag{115}$$

and the signal-to-noise power ratio at its output is

$$\rho_{max} = \int_{-\infty}^{\infty} \frac{P_s(s)}{P_n(s)}\, ds \tag{116}$$

First, notice that while $H_o(s)$ is real and even, hence containing no phase information, $K_o(s)$ is Hermite and does contain phase information. Notice also that $H_o(s)$ is bounded between 0 and +1. This means it can never amplify spectral components of the input signal. However $K_o(s)$ has neither positive nor negative bound. Therefore its frequency domain behavior is much less constrained.

Let us define the signal-to-noise power ratio as a function of frequency by

$$R(s) = \frac{|S(s)|^2}{|N(s)|^2} = \frac{P_s(s)}{P_n(s)} \tag{117}$$

In terms of this function, the magnitude of the matched detector transfer function is given by

$$|K_o(s)| = \frac{R(s)}{|S(s)|} = \frac{\sqrt{R(s)}}{|N(s)|} \tag{118}$$

and the signal-to-noise ratio by

$$\rho_{max} = \int_{-\infty}^{\infty} R(s)\, ds \tag{119}$$

The Wiener filter transfer function is

$$|H_o(s)| = H_o(s) = \frac{R(s)}{1 + R(s)} \tag{120}$$

and the mean-square error is given by

$$\text{MSE}_o = \int_{-\infty}^{\infty} \frac{R(s)P_n(s)}{1 + R(s)} \tag{121}$$

Comparing Eqs. (120) and (121), we see that

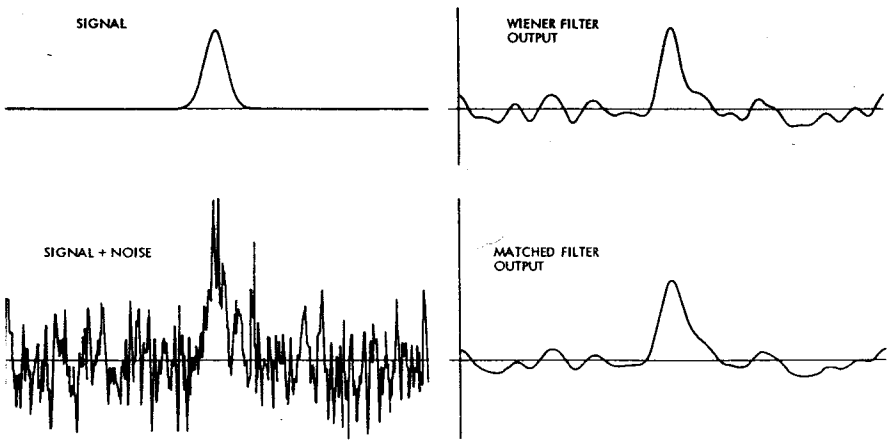
$$\text{MSE}_o = \int_{-\infty}^{\infty} P_n(s)H_o(s)\,ds \tag{122}$$

which indicates that the mean-square error is just the noise power that passes through the filter accumulated over all frequencies.

In a sense, estimation is a more difficult task than detection. There are two reasons for this. First, we ask an estimator to recover the signal at all points in time, whereas we ask the detector only to determine when the signal occurs. Second, we have more *a priori* information in a detection problem in that we know the form of the signal exactly instead of having only its power spectrum. Since we are asking a detector to do less with more information, we can expect better performance under the same conditions.

Whether one uses a detector or an estimator is dictated by the problem. Since they are designed for different jobs, they usually do not compete for consideration. Nevertheless, it is instructive to compare the behavior of the two filters under similar conditions. Figure 11-21 presents a computer simulation that illustrates both the Wiener estimator and the matched detector when the signal is a Gaussian pulse embedded in white random noise. In this case, the signal-to-noise ratio is on the order of unity.

Both the estimator [Eq. (113)] and the detector [Eq. (115)] are lowpass filters in




igure 11-21 Comparison of the Wiener and matched filters



this situation, but they differ somewhat in form. The detector output clearly shows a peak at the point where the input pulse occurs. The estimator recovers the pulse from the noise, but not without residual error. The low-frequency components of the noise penetrate the Wiener filter and prevent exact recovery. One would expect better performance from both filters with improved signal-to-noise ratio, and conversely.

### A Practical Example

We conclude this chapter with an example that illustrates how optimal filter theory can guide the design of practical filters. Figure 11-22 shows a digitized X ray of a tube filled with X-ray absorbing dye. This models angiography, the diagnostic technique in

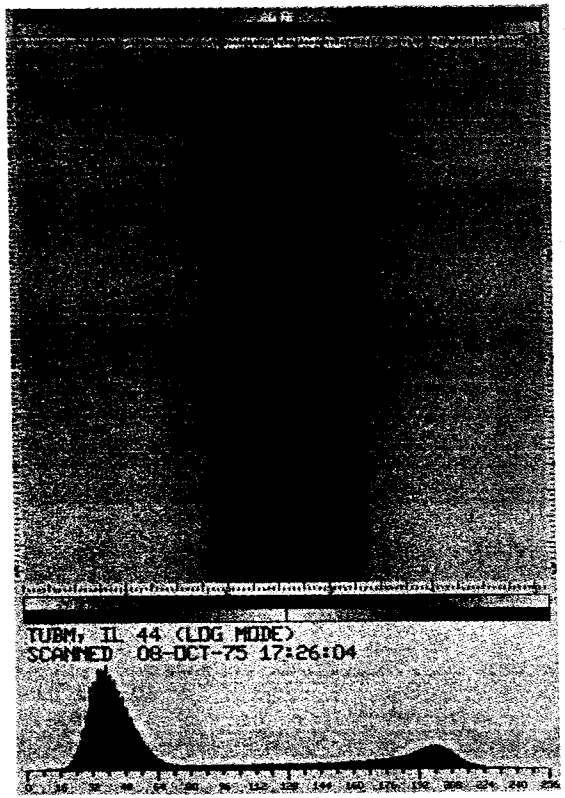



**Figure 11-22** Digitized angiogram of a smooth tube

which dye is injected into blood vessels during X-ray exposure. Here the smooth tube substitutes for the vessel. The goal in this example is a processing technique that will find the tube edges in the noisy image of Figure 11-22 and reliably measure tube diameter all along its length. Such a technique is useful for quantifying the blood vessel narrowing which accompanies atherosclerosis and produces heart attacks (Ref. 7).

Since this an edge detection problem, the matched detector seems a natural choice. In this example, however, we shall pose the problem somewhat differently. We shall assume the vessel edges occur, on each image line, at the two points of steepest slope, and locate those by differentiation. Before differentiating, however, we shall use a Wiener filter to estimate the noise-free image. Furthermore, we shall process each line individually so that the procedure can respond to rapid changes in width, should they occur.

Figure 11-23 shows a gray level plot of one line $f_i(x)$ from Figure 11-22. The evident noise is common in radiography, due primarily to film grain and photon statistics in the illuminating beam. Clearly, differentiating this curve would not produce reliable peaks at the inflection points.

Assuming uncorrelated signal $s(x)$ and noise $n(x)$, the specification of the Wiener filter [Eq. (58)] requires the power spectrum of the signal and that of the noise. We can estimate the signal power spectrum by line averaging since, with a smooth tube, all lines $f_i(x)$ should be identical in the absence of noise. Thus

$$P_s(s) = |\mathfrak{F}\{s(x)\}|^2 \approx \left|\mathfrak{F}\left\{\frac{1}{N}\sum_{i=1}^{N} f_i(x)\right\}\right|^2 \tag{123}$$

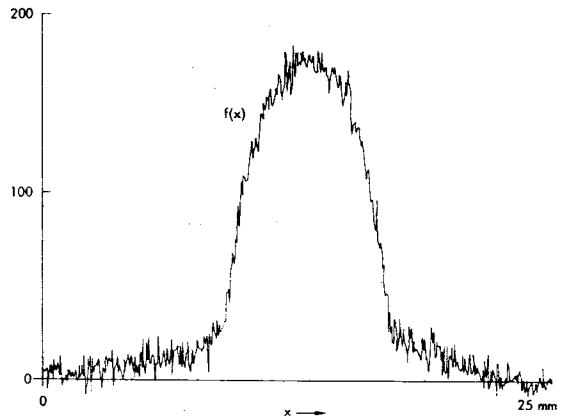



**Figure 11-23** Line 100 from Figure 11-22



will reduce the noise by the factor $1/\sqrt{N}$. Figure 11-24 shows the result of averaging 60 lines in Figure 11-22, and the resulting signal amplitude spectrum.

Once the signal has been estimated, the noise power spectrum can be estimated from Figure 11-22 using line by line power spectrum averaging after signal subtraction, that is

$$P_n(s) \approx \frac{1}{N}\sum_{i=1}^{N} |\mathfrak{F}\{f_i(x) - s(x)\}|^2 \tag{124}$$

In this study Eq. (124) showed the noise power spectrum to be essentially constant with frequency (white noise).

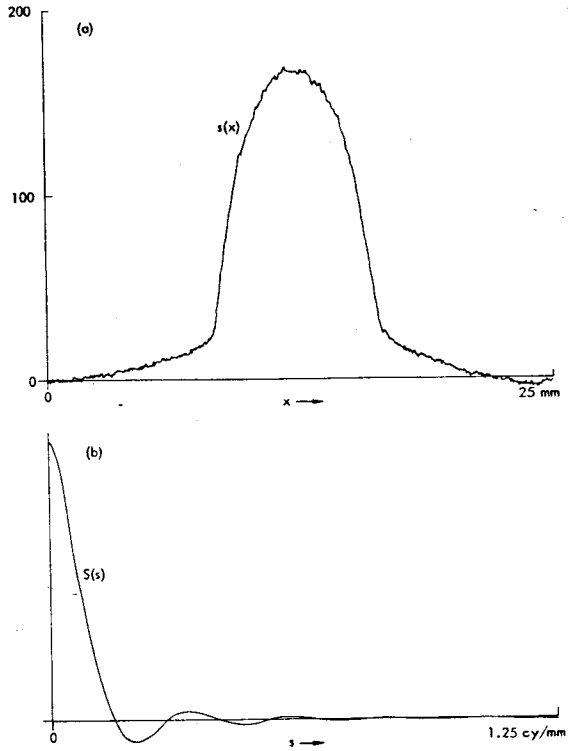



**Figure 11-24** (a) Noise-free signal estimate obtained by line averaging in Figure 11-22; (b) Fourier amplitude spectrum of (a)

Figure 11-25(a) shows the Wiener filter transfer function $H_o(s)$ computed by Eq. (58). The transfer function takes on values near unity at the signal dominated low frequencies, and tends to zero at high frequencies. We could inverse transform the transfer function in Figure 11-25(a) to obtain the impulse response for pre-differentiation smoothing. There are, however, some practical considerations worthy of note.

The notches in the transfer function of Figure 11-25(a) are produced by the zero crossings in the signal spectrum [Figure 11-24(b)]. By the similarity theorem, the position of these notches will shift with changes in vessel width. This points out that our signal is not actually an ergodic random process as the Wiener filter development assumes. The member functions in the signal ensemble correspond to vessels of different width and thus do not all have identical power spectra. As it happens, we are forced to violate one of the assumptions on which the Wiener filter is based. We shall nevertheless proceed, acting in the belief that a "near-optimal" technique will prove an adequate substitute for optimality.

If we include the troublesome notches, our filter will be quite sensitive to slight changes in vessel width. We choose instead to ignore the notches by fitting a smooth envelope to the transfer function. Figure 11-25(b) shows a smooth approximation $\tilde{H}(s)$ to the Wiener filter transfer function. This function was chosen because of two desirable properties. First, it is a reasonable approximation to the envelope of Figure

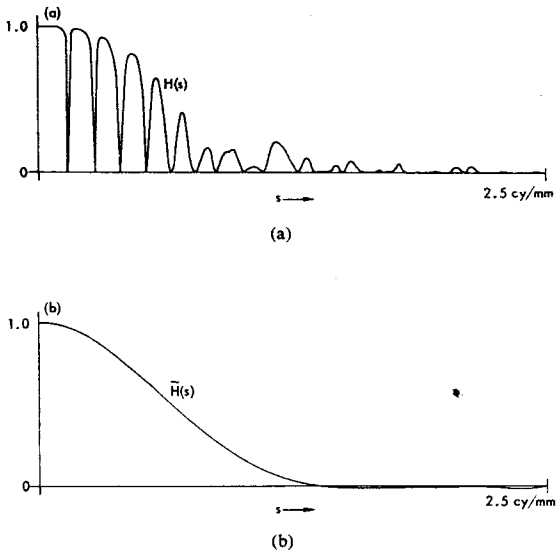



**Figure 11-25** (a) Wiener filter transfer function; (b) Smooth approximation to (a)



11-25(a). Second, its impulse response renders digital convolution quite an efficient computation.

Figure 11-26 shows the corresponding impulse response $\tilde{h}(x)$, which is piecewise parabolic, and its piecewise linear first derivative $\tilde{h}'(x)$. Since differentiation commutes with convolution, using the latter function combines smoothing and differentiation into one step. Furthermore, digital convolution using a piecewise linear impulse response can be programmed to execute very efficiently (Ref. 8).

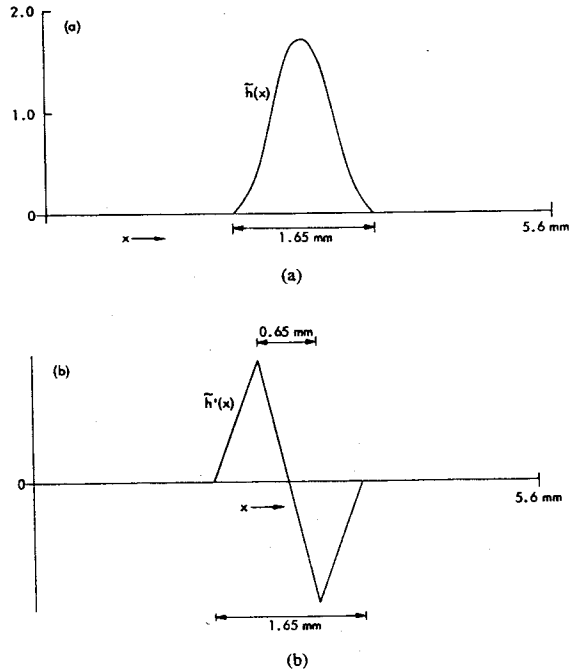



**Figure 11-26** (a) Impulse response of Figure 11-25(b); (b) Derivative of (a)

Figure 11-27 shows the results of using the two impulse responses in Figure 11-26 on the image line in Figure 11-23. The first produces smoothing for noise reduction only while the second combines smoothing with differentiation. In this case the degree of noise reduction is gratifying. Notice also that the inflection points in the upper curve give rise to distinct peaks in the lower curve, suggesting that vessel edge detection is now a simple task.

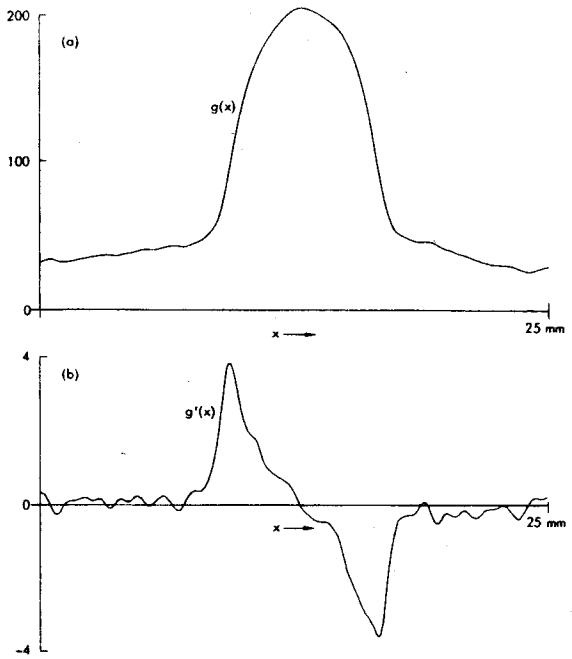



**Figure 11-27** (a) Result of smoothing the line in Figure 11-23 with $\tilde{h}(x)$; (b) with $\hat{h}'(x)$

The piecewise linear impulse response $\hat{h}'(x)$ is a computationally efficient approximation to the differentiating Wiener filter for this application. Even though our signal is non-ergodic, the notch-free transfer function $\hat{H}(s)$ should be rather well-behaved under suboptimal conditions since it has no abrupt behavior in the frequency domain. Furthermore, Figure 11-27 strongly suggests that we have a comfortable solution to this edge detection problem. The differentiating Wiener filter designed on the smooth tube has proved useful on routine angiograms (Ref. 8).

## SUMMARY OF IMPORTANT POINTS

1. A high-frequency enhancement filter impulse response can be designed as the sum of a narrow positive pulse and a broad negative pulse.
2. The transfer function of such a filter approaches a maximum value that is equal to the area under the narrow positive pulse.



3. The transfer function of such a filter has a zero-frequency response equal to the difference of the areas under the two component pulses.
4. The zero-frequency response of a filter determines how the contrast of large features is affected.
5. Filters designed for ease of computation rather than for optimal performance are likely to introduce artifacts into an image.
6. An ergodic random process is a signal whose known power spectrum and autocorrelation function represent all the available knowledge.
7. The Wiener estimator is optimal, in the mean-square error sense, for recovering a signal of known power spectrum from additive noise.
8. The Wiener filter transfer function takes on values near unity in frequency bands of high signal-to-noise ratio and near zero in bands dominated by noise.
9. The matched detector is optimal for detecting the occurrence of a known signal in a background of additive noise.
10. In the case of white noise, the matched filter correlates the input with the known signal.
11. The Wiener filter transfer function is real and even and bounded by zero and unity.
12. The matched filter transfer function is complex and Hermite and in general unbounded.

## REFERENCES

1. N. WIENER, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, John Wiley & Sons, New York, 1949.
2. W. B. DAVENPORT and W. L. ROOT, *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill Book Company, New York, 1958.
3. Y. W. LEE, *Statistical Theory of Communication*, John Wiley & Sons, New York, 1960.
4. L. A. WAINSTEIN and V. D. ZUBAKOV, *Extraction of Signals From Noise*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962.
5. G. L. TURIN, "An Introduction to Matched Filters," *IRE Transactions on Information Theory*, 311–329, June 1960.
6. D. MIDDLETON, "On New Classes of Matched Filters and Generalizations of the Matched Filter Concept," *IRE Transactions on Information Theory*, 349–360, June 1960.
7. E. S. BECKENBACH, R. H. SELZER, D. W. CRAWFORD, S. H. BROOKS, and D. H. BLANKENHORN, "Computer Tracking and Measurement of Blood Vessel Shadows from Arteriograms," *Medical Instrumentation*, **8**, No. 5, September–October, 1974.
8. K. R. CASTLEMAN, R. H. SELZER, and D. H. BLANKENHORN, "Vessel Edge Detection in Angiograms: An Application of the Wiener Filter," in J. K. AGGARWAL, ed., *Digital Signal Processing*, Point Lobos Press, 13000 Raymer St., No. Hollywood, California 91605, 1979.