Handbook of Measurement Science, Vol. 1
Edited by P. H. Sydenham


Chapter

# 11

D. M. MUNROE

# Signal-to-Noise Ratio Improvement

*Editorial introduction*

A measurement system begins with sensing stages that couple to relevant measurands of the system under study. The power level of the information-bearing signals formed by the sensors is often very low and may be swamped by the unwanted noise signals that are present. Careful attention to sensor and circuit design and assembly, plus use of certain signal processing methods, makes it possible to greatly enhance the original signal-to-noise ratio to usable levels.

This chapter discusses the various strategies that are available providing a basis for their use (the best information is available in the literature of companies marketing such products). It is surprising to find that, after well over half a century of analog signal processing progress, there still exists no full length text that deals generally with signal recovery and enhancement in general instrumentation applications. This chapter is unique in this respect, apparently being the first to appear in a published text. It deals with this material at an extended theoretical depth.

It will be noticed that techniques, originally conceived for analog signals, are gradually being implemented in digital form, thus in many cases providing improved performance at comparable or less cost. This trend can be expected to increase with time as digitally oriented designers, seeking improved means of signal recovery, become more familiar with the principles already proved in the analog signal domain.

## 11.1 INTRODUCTION

Recovering or enhancing a signal or improving a signal-to-noise ratio simply means reducing the noise accompanying a signal. There are two basic ways of doing this:

(a) Bandwidth reduction, where the noise is reduced by reducing the system noise bandwidth ($B_n$). This approach works well if the frequency spectra of the noise and signal do not overlap significantly, so that reducing the noise bandwidth does not affect the signal. With random white noise the output noise is proportional to $\sqrt{B_n}$.

(b) Averaging or integrating techniques, where successive samples of the signal are synchronized and added together. The signal will grow as the number ($n$) of added samples; with random white noise the noise will grow as $\sqrt{n}$.

In many applications there is significant overlap between the signal and noise spectra and improving a signal-to-noise ratio must be done at the expense of the response time or measurement time ($T$); with random white noise interference the output signal-to-noise ratio is proportional to $\sqrt{T}$. The bandwidth reduction technique is best looked at from a frequency-domain point of view; signal averaging and correlation techniques lend themselves to time-domain analysis.

In this chapter, mathematics and theoretical considerations will be reduced to a minimum; the reader is referred to Chapter 4 for additional theoretical and background information. For further simplicity we will assume that all noise processes are stationary and that both signal and noise are ergodic, *analog* variables; we will not concern ourselves with digital signals or discrete-time (sampled) signals except where such signals are involved in the enhancement techniques. In addition, only signal recovery techniques will be considered. Further processing, such as least-squares polynomial smoothing of a waveform or Fourier transformation to obtain a frequency spectrum, will not be considered here.

We will start by reviewing some basic concepts, move on to discuss ways to avoid *adding* noise (e.g. hum pick-up and preamplifier noise) and then discuss instrumentational techniques to reduce the remaining noise content. Finally, we will discuss some of the special considerations involved in recovering pulse signals from photon (light), ion, or electron beams.

## 11.2 NOISE AND NOISE BANDWIDTH

Noise is an undesired signal. It usually becomes of interest when it obscures a desired signal. Figure 11.1 shows the power spectral density (power/unit bandwidth) of the most commonly encountered types of noise.

Deterministic noise can range from simple discrete-frequency components such as power-line hum at harmonics of 50 or 60 Hz, to wide-band interference (RFI) caused by narrow, high-energy pulses from power-line switching spikes, pulsed lasers, radar transmitters, and the like.

Stochastic (random) noise is found in most systems both as white noise, where the power spectral density is independent of frequency, and also as $1/f$ or flicker noise, where the power spectral density decreases as frequency increases. Power spectral density is usually measured in mean-squared-volts/Hz or mean-squared-amperes/Hz; for noise, such specifications are usually referred to as *spot noise* data and usually are a function of frequency. Notice

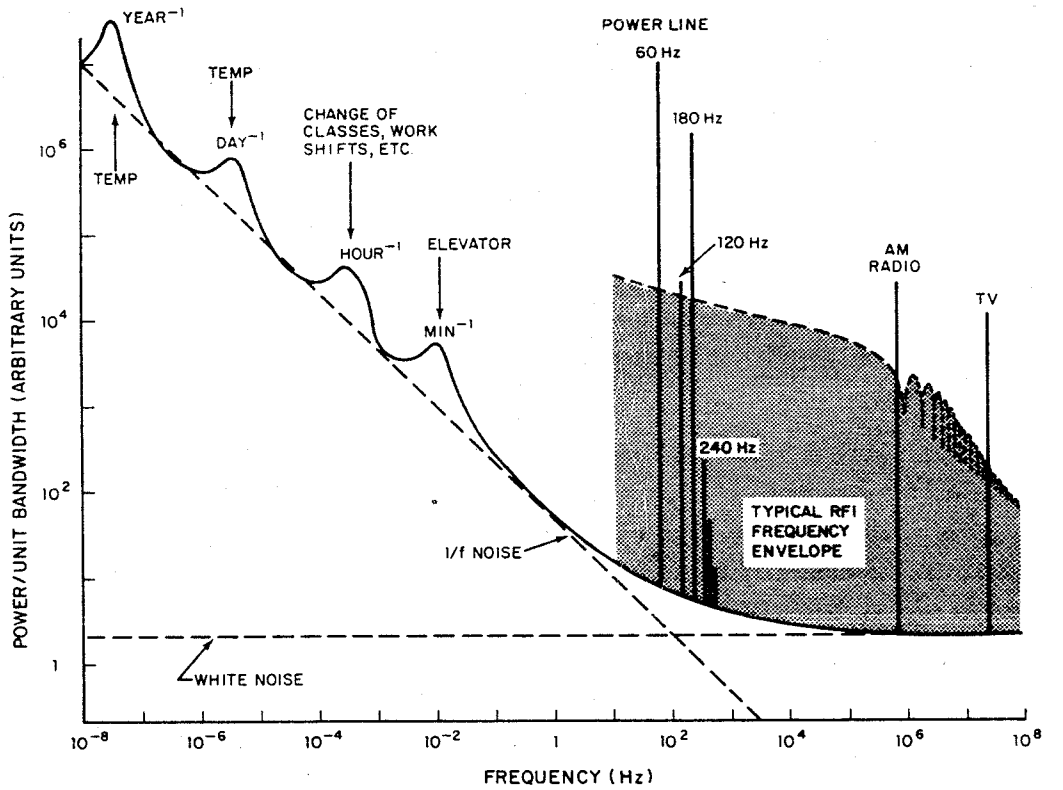



Figure 11.1 Environmental noise (reproduced by permission of EG&G Princeton Applied Research Corporation)

that for an r.m.s. voltage of $e$ (volts) and a frequency range of $\Delta f$ (Hz), the power spectral density, $S$, is given by

$$S = \frac{e^2}{\Delta f} = \left(\frac{e}{\sqrt{(\Delta f)}}\right)^2 \tag{11.1}$$

The quantity $e/\sqrt{(\Delta f)}$ is usually referred to as *voltage spectral density* and is measured in r.m.s.-volts/$\sqrt{\text{Hz}}$ (volts per root hertz). Similarly, we can refer to current spectral density specifications in units of r.m.s.-amperes/$\sqrt{\text{Hz}}$.

White noise is usually found in one of two forms: *Johnson* noise and *shot* noise. Johnson, or thermal, noise is caused by random motion of thermally agitated electrons in resistive materials, and the mean-square noise voltage is given by

$$e_n^2 = 4kTR\,\Delta f \tag{11.2}$$

where $k$ is Boltzmann's constant ($1.381 \times 10^{-23}$ JK$^{-1}$), $T$ is the absolute temperature (kelvin) and $R$ is the resistance (ohm). Alternatively, from Ohm's law, the mean-square noise *current* is given by

$$i_n^2 = \left(\frac{e_n}{R}\right)^2 = \frac{4kT\Delta f}{R} \tag{11.3}$$

Shot noise is caused by the random arrival of electrons (see Section 11.10.2) at, for example, the electrodes of electron tubes or transistor junctions. A d.c. current, $I$, will have a noise-current component, $i_n$, given by

$$i_n^2 = 2AeI\Delta f \tag{11.4a}$$

where $e$ is the charge of one electron ($\simeq 1.6 \times 10^{-19}$ C), $A$ is the mean gain experienced by each electron and $I$ is in amperes. In many cases (see Section 11.10.2), $A = 1$, so that

$$i_n^2 = 2eI\Delta f \tag{11.4b}$$

Flicker noise has many different origins and is not clearly understood but exhibits a $1/f^n$ power spectrum with $n$ usually in the range 0.9 to 1.35. Note that *d.c. drift* is a very-low-frequency form of flicker noise.

What do we mean by *bandwidth*? In the simple low-pass filter circuit shown in Figure 11.2a, for example, we usually and somewhat arbitrarily define the *signal bandwidth* (Figure 11.2b) to be the *cut-off frequency*, $f_c$, where $e_o/e_i = 70.7\%$ (−3 dB) or $e_o^2/e_i^2 = 50\%$ (the half-power point).

Notice that frequencies above $f_c$ will obviously pass (though attenuated) through the filter, and therefore are not really cut off. For noise, it is convenient

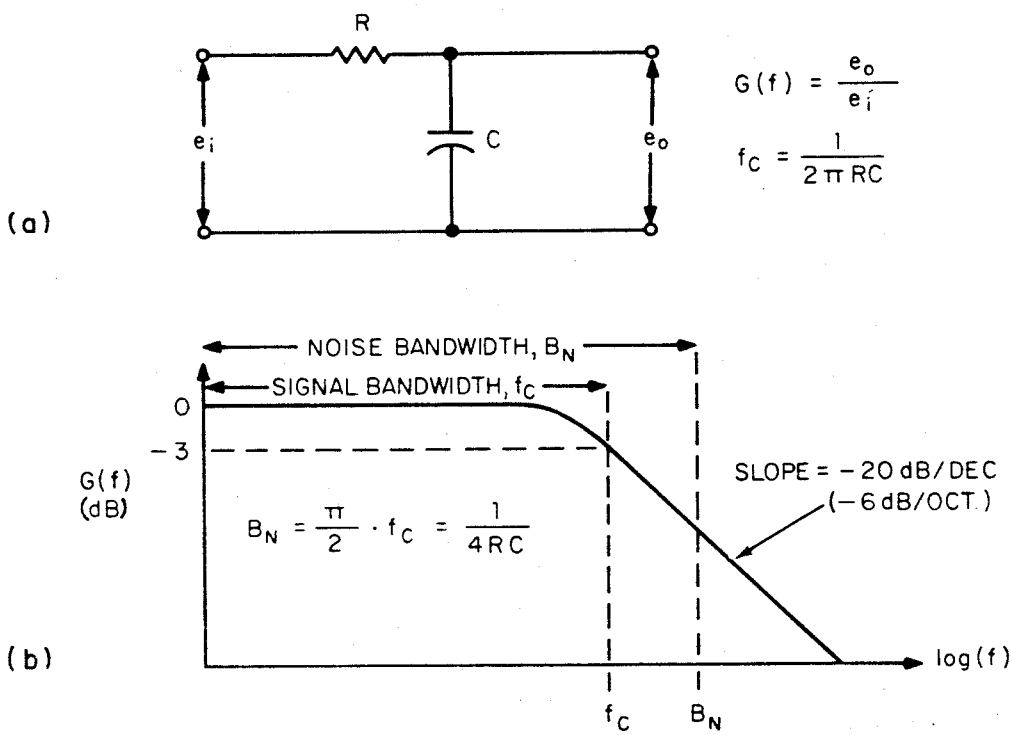



Figure 11.2 Signal and noise bandwidths of low-pass filter: (a) circuit; (b) Bode plot

to think in terms of an equivalent noise bandwidth, $B_n$, defined by the relationship

$$B_n = \frac{1}{G^2} \int_0^\infty |H(j\omega)|^2 \, df \tag{11.5}$$

where $H(j\omega)$ is the frequency response function of the system and $G$ is a gain parameter suitably chosen to be a measure of the response of the system to some parameter of the signal: for low-pass systems (e.g. Figure 11.2b) $G$ is usually taken to be the zero-frequency (d.c.) gain; for band-pass responses, $G$ is usually made equal to the maximum gain.

Using the above definition, and taking $G$ to be the zero-frequency gain (i.e. unity), we can readily calculate that for the simple $RC$ filter shown in Figure 11.2a:

$$B_n = 1/4RC \text{ Hz} \tag{11.6}$$

Noise, of the stochastic form, is reviewed in relation to instrument systems in Fellgett and Usher (1980).

## 11.3 SIGNALS AND SIGNAL-TO-NOISE RATIO

Suppose we were to look at a complex waveform on an oscilloscope. What is the signal? Is it the complete waveform? The peak (or r.m.s. or average) amplitude? The depth of modulation? The implied frequency spectrum? The difference in time or amplitude between two features of the waveform? The answer, of course, is that the information-bearing signal could be any or none of the above. In this chapter, we will restrict ourselves to some commonly encountered types of signal where enhancement is often required. Together with the enhancement technique normally used, these are:

(a) base-band (d.c.) signals: low-pass filtering or autocorrelation;
(b) amplitude modulated signals: band-pass filtering or phase-sensitive detection;
(c) repetitive (not necessarily periodic) swept signals: signal averagers;
(d) photon, electron or ion beam signals: photon-counting systems.

The word *signal* is often used rather ambiguously to mean either the total signal being measured or a noise-free, information-bearing component of it. The following definitions should allow us to avoid such confusion. We will normally talk in terms of a *total signal* consisting of an r.m.s. signal component ($S$) accompanied by an r.m.s. noise component ($N$). Thus

$$\text{Signal-to-noise ratio, SNR} = S/N \tag{11.7}$$

Note that

$$\text{measurement uncertainty or inaccuracy} = \frac{1}{\text{SNR}} \tag{11.8}$$

$$\text{Signal-to-Noise improvement ratio (SNIR)} = \frac{\text{SNR}_{\text{out}}}{\text{SNR}_{\text{in}}} = \frac{S_o/N_o}{S_i/N_i} \tag{11.9}$$

For unity gain (i.e. $S_o = S_i$), band-limited white input noise of bandwidth $B_{ni}$ and output noise bandwidth $B_{no}$,

$$\text{SNIR} = N_o/N_i = \sqrt{(B_{ni}/B_{no})} \tag{11.10}$$

## 11.4 NOISE MATCHING AND PREAMPLIFIER SELECTION

All preamplifiers add noise. Whether this additional noise is significant will depend, of course, upon the noise level from the signal source. Since uncorrelated noise adds vectorially (in an r.m.s. fashion), the preamplifier noise can be neglected if it is less than about one-third of the source noise:

$$\sqrt{[(1.0)^2 + (0.3)^2]} \simeq 1.0$$

We can think of a practical preamplifier as consisting of an ideal, noise-free amplifier with a (frequency-dependent) noise-voltage generator of voltage spectral density $e_n$ (V/$\sqrt{}$Hz), and a noise-current generator of current spectral density $i_n$ (A/$\sqrt{}$Hz), connected to its input as shown in Figure 11.3a. Figures 11.3b and 11.3c, respectively, show separately the gain seen by the amplifier internal noise voltage and current generators. Any input shunt capacitance (see Figure 11.3b) will decrease the input impedance and cause output noise that increases with frequency if $Z_f$ is resistive.

The preamplifier noise may also be defined (Faulkner, 1966) in terms of an equivalent series noise resistance $R_e$, and an equivalent parallel noise resistance, $R_i$, where (from equation (11.2))

$$R_e = \frac{e_n^2}{4kT\Delta f} \text{ ohms}$$

and (from equation (11.3))

$$R_i = \frac{4kT\Delta f}{i_n^2} \text{ ohms}$$

We can also define the *noise figure* (NF) of the preamplifier to be (in dB)

$$\text{NF} = 10 \log_{10}\left(1 + \frac{R_e}{R_s} + \frac{R_s}{R_i}\right) \tag{11.11}$$

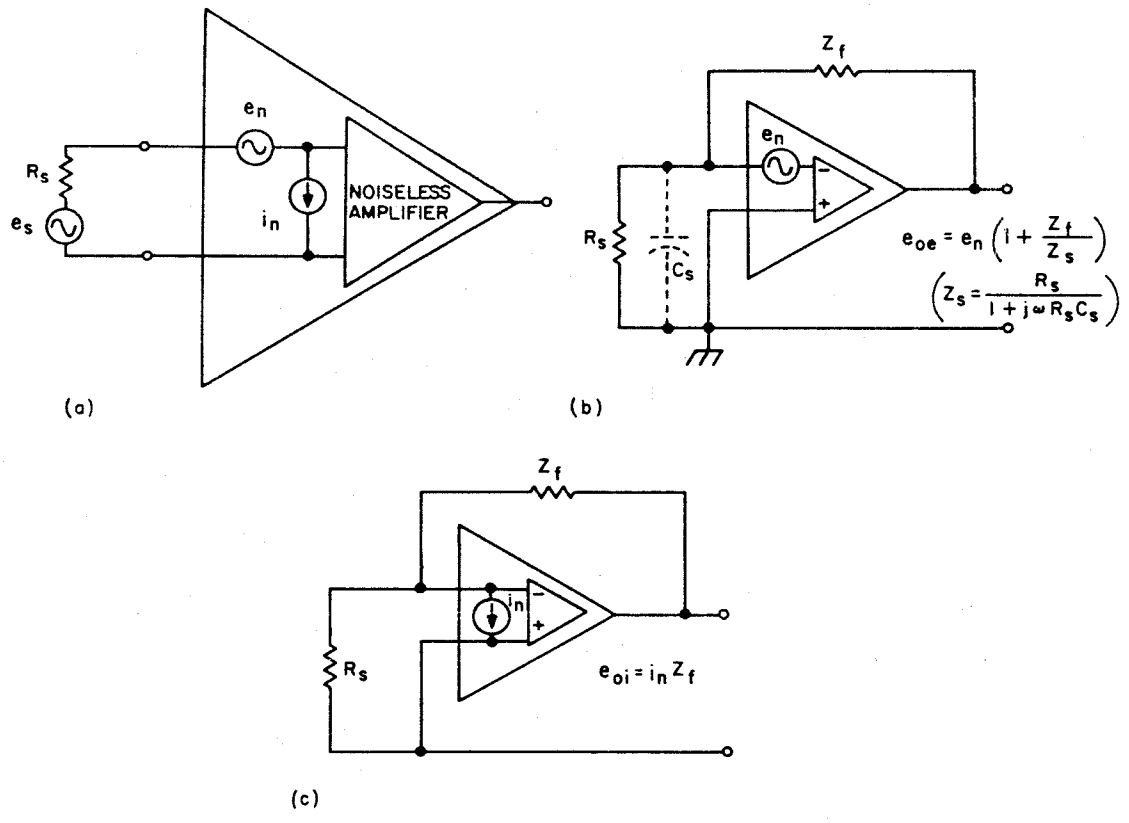



Figure 11.3 Amplifier noise: (a) equivalent circuit; (b) voltage noise; (c) current noise

A perfect or noiseless preamplifier would have a 0 dB noise figure. Figure 11.4 shows the noise figure contours that result when the noise figure of a practical preamplifier is plotted as a function of source resistance and frequency. Notice from equation (11.11), that with high source resistance, $R_e/R_s \to 0$, and

$$NF \simeq 10 \log_{10}\left(1 + \frac{R_s}{R_i}\right)$$

and the amplifier noise current, $i_n$, predominates. With low source resistances, the amplifier noise voltage, $e_n$, becomes the major noise source. Wherever possible, preamplifiers should be chosen so that their 3 dB noise-figure contour encloses the expected range of source resistance and frequency.

For a given preamplifier, the optimum source resistance, $R_s$, is given by

$$R_s(\text{opt}) = \frac{e_n}{i_n} = \sqrt{(R_e R_i)} \text{ ohms} \qquad (11.12)$$

Note that adding a series or parallel resistance between the signal source and the preamplifier always reduces signal and adds noise, and so cannot be used to obtain a *better match.*

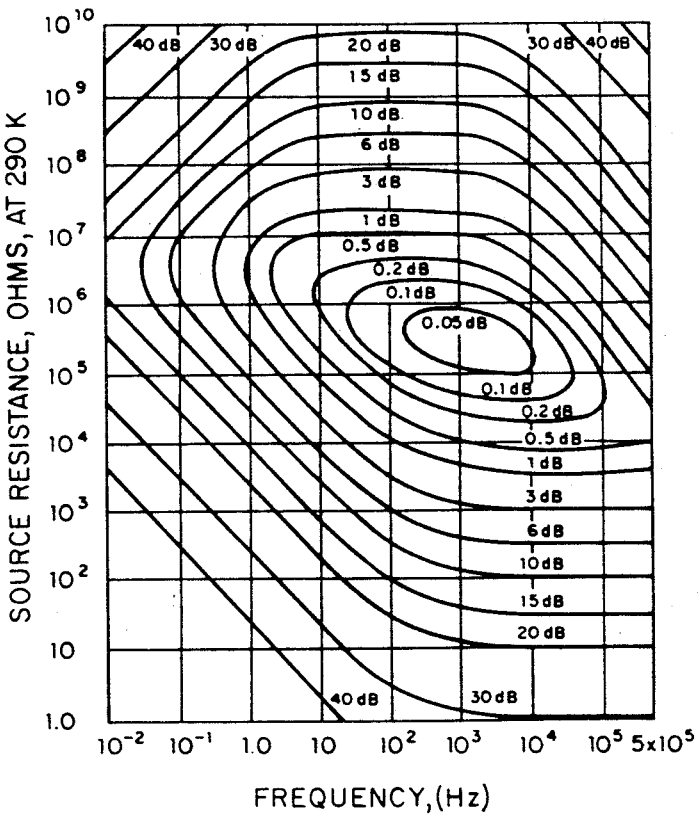



Figure 11.4 Typical noise figure contours for a high input impedance preamplifier (reproduced by permission of EG&G Princeton Applied Research Corporation)

Preamplifiers can be classified in many ways; one basic division, for example, is between *differential* input and *single-ended* input. All other things being equal, a differential preamplifier generates 3 dB (41.4%) more noise than a single-ended version. However, this disadvantage is significant only in situations where preamplifier noise predominates and, in many cases, is outweighed by the flexibility of a differential input and its ability to remove ground-loop problems (see Section 11.5).

Transformers are often used to match very low source impedances (0.1 Ω–1 kΩ). Figure 11.5 shows an amplifier with an optimum, $\sqrt{(R_e R_i)}$, source resistance value of 1 MΩ being matched to a 100 Ω thermopile by means of a 100 : 1 voltage step-up transformer (10,000 : 1 impedance transformation). Note that, in general, such noise matching does not result in the same circuit values as would power matching; that is, the amplifier input resistance is not normally equal to $\sqrt{(R_e R_i)}$. Transformers should be avoided if possible, since they reduce frequency response, may pick up magnetically induced interference and may be microphonic.

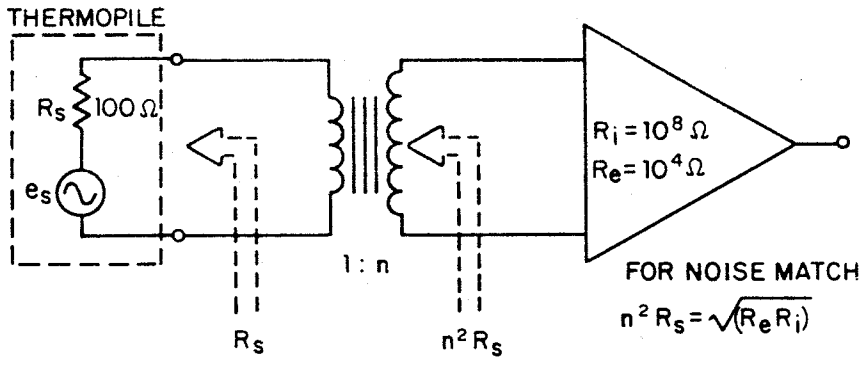



Figure 11.5 Transformer noise matching

For sources of approximately 100 Ω–10 kΩ, preamplifiers are available which use an input stage consisting of multiple bipolar transistors connected in parallel to provide a lower value of $\sqrt{(R_e R_i)}$. Such preamplifiers avoid the bandwidth constraints imposed by input transformers. For higher impedance sources (1 kΩ–100 MΩ), preamplifiers usually employ junction-FET's as input devices and are available as voltage preamplifiers, charge amplifiers (for use with capacitive transducers), or current-input (transresistance) amplifiers. (See Figure 11.6).

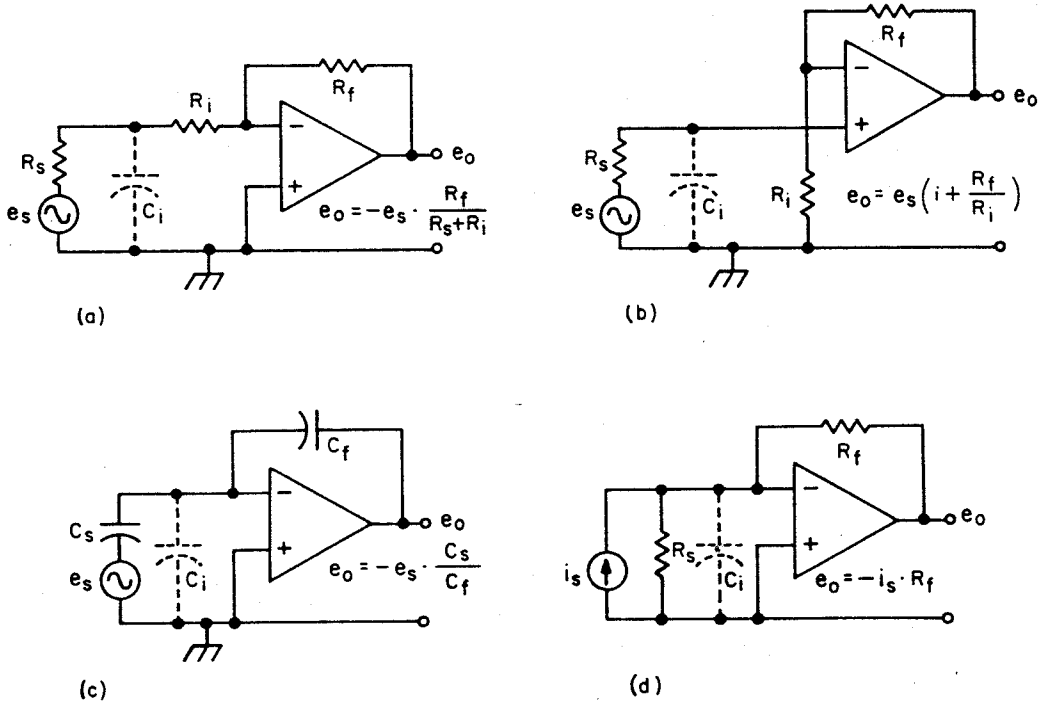



Figure 11.6 Amplifier configurations: (a) voltage, inverting; (b) voltage, non-inverting; (c) charge; (d) current input (trans-resistance)

In Figures 11.6a and 11.6b any cable capacitance or stray capacitance, $C_i$, will form a low-pass filter with the source resistance, $R_s$, having a $-3$ dB frequency given by

$$f_c = \frac{1}{2\pi R_i R_s C_i/(R_i + R_s)} \quad \text{(Figure 11.6a)}$$

or

$$f_c = \frac{1}{2\pi R_s C_i} \quad \text{(Figure 11.6b)}$$

In Figures 11.6c and 11.6d, such shunt capacitance appears at first sight to have no effect since it is effectively short-circuited by the virtual-ground input. However, as shown in Figure 11.3b, shunt capacitance will cause a deterioration in the output SNR and also (by introducing an additional pole into the loop gain) may cause *ringing* in the amplifier response, or even oscillation. By careful design (which usually includes adding a capacitor across the feedback resistor) these effects can be minimized and with high source impedance, commercial current and charge amplifiers usually provide significantly greater bandwidth than can corresponding voltage amplifiers.

## 11.5 INPUT CONNECTIONS; GROUNDING AND SHIELDING

Ideally, all grounds should have a zero-impedance connection to each other and to wet earth; in practice they do not. Due to voltage drops across their finite impedance to earth, capacitively or inductively coupled interference, and other reasons, each ground tends to be at a different potential from other nearby grounds. If two (or more) such adjacent grounds are connected together to form a ground loop (Figure 11.7a), then the potential difference between the grounds will cause a circulating current. The potential difference between grounds ($e_{cm}$), is called the *common-mode source* since it is common to both the signal (via loop 2) and ground (via loop 1) inputs of the preamplifier.

Figure 11.7b rearranges the circuit of Figure 11.7a and assumes the signal source, $e_s$, to be zero. Note that the low resistance of the coaxial cable shield (braid), $R_{cg}$, is in parallel with the series combination of the source resistance, $R_s$, the coaxial cable centre conductor resistance $R_{cs}$ and the preamplifier input impedance, $Z_{in}$. Under normal circumstances $(R_s + R_{cs} + Z_{in}) \gg R_{cg}$ and $Z_{in} \gg (R_s + R_{cs})$, so that as shown in Figure 11.7c, the common-mode voltage dropped across $R_{cg}$ is also applied across the preamplifier input terminals. More generally, with $e_s = 0$, the preamplifier input voltage, $e_{in}$, is given by

$$e_{in} = e_{cm} \frac{R_{cg}}{R_{cg} + R_{sg} + R_{pg}} \tag{11.13}$$

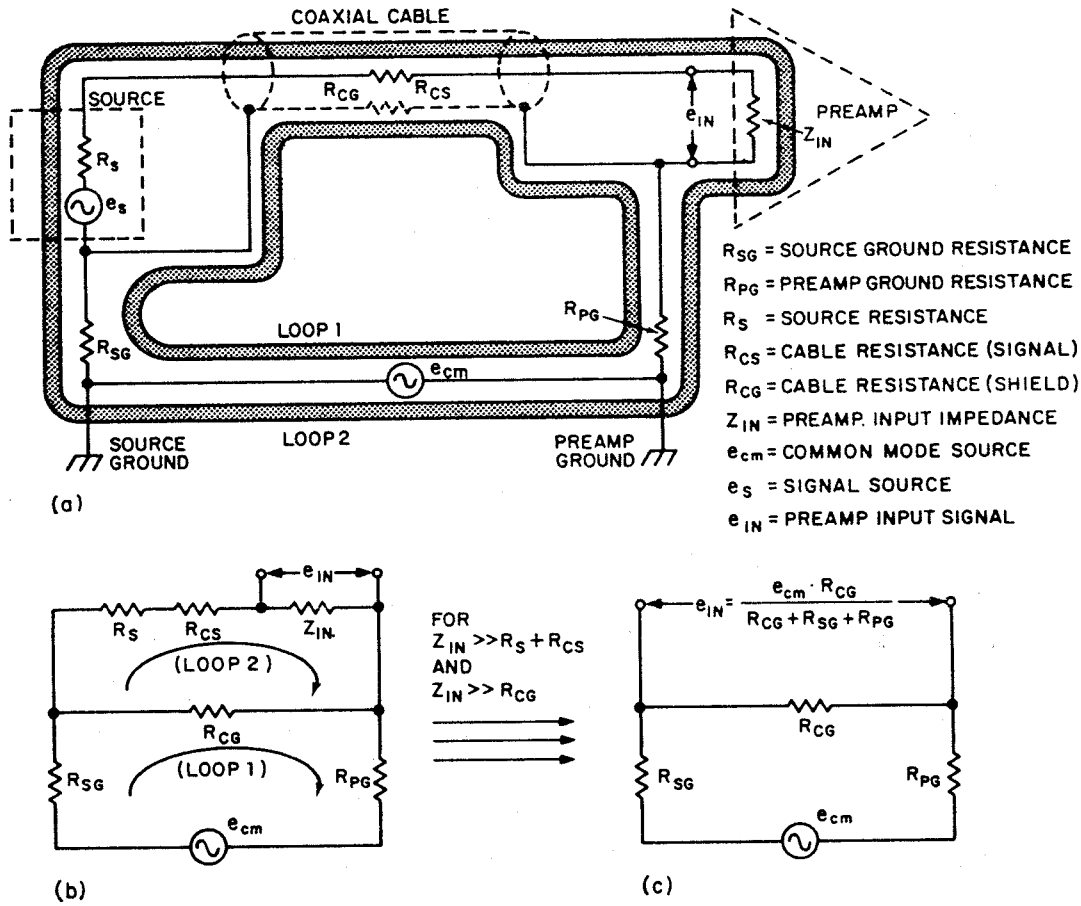



Figure 11.7 Ground loops: (a) physical occurrence; (b) schematic equivalent circuit; (c) reduced equivalent circuit

From equation (11.13), this common-mode input to the preamplifier can be removed (i.e. $e_{in} = 0$) by making:

(a) $e_{cm} = 0$. This can be attempted by grounding the source and preamplifier to the same ground point, and shielding to remove capacitively or inductively coupled interference, but the procedure is rarely completely successful.

(b) $R_{cg} = 0$. The usual approach here is to bolt both source and preamplifier chassis to a large metal plate. Unfortunately, it is fairly easy to develop large potential differences between points a centimetre or two apart on a large metal plate, such as a mounting rack.

(c) $R_{sg} = \infty$. *Floating* or disconnecting the source from ground is a good approach where practicable.

(d) $R_{pg} = \infty$. The preamplifier may also be floated—particularly if it is battery powered. Note that disconnecting the power-line ground from an instrument can be extremely dangerous. In many instruments $R_{pg}$ consists of an internal 10 Ω–1 kΩ resistor that can be switched into the circuit to effectively float the amplifier input terminals.

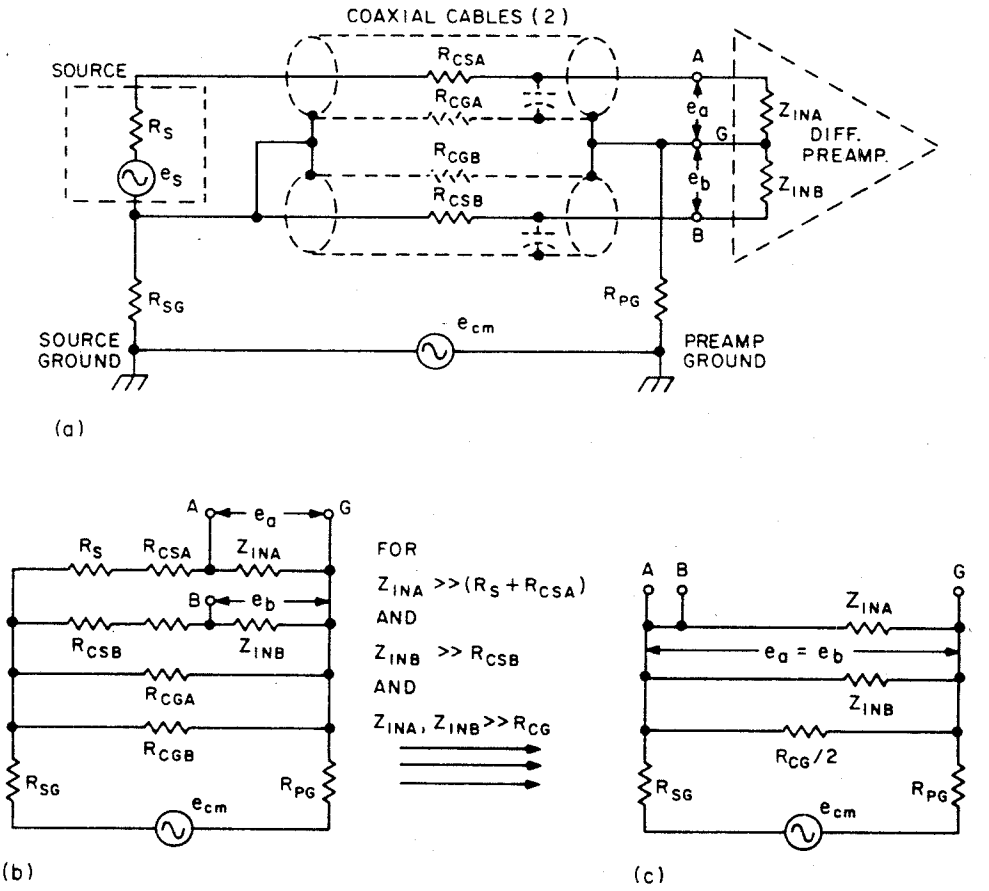



Figure 11.8 Differential preamplifier used with single-ended source: (a) physical occurrence; (b) schematic equivalent circuit; (c) reduced equivalent circuit

Figure 11.8 illustrates the use of a differential amplifier with an unbalanced (single-ended) source to eliminate or reduce ground-loop problems. As in Figure 11.7, the circuit simplification assumes that the input impedance of each side of the differential amplifier ($Z_{inA}$ and $Z_{inB}$) is much larger than source or cable resistances. At low frequencies this differential connection results in equal common mode voltages at the amplifier's input terminals (A and B), and the amplifier's ability to discriminate against common-mode signals (i.e. its *common-mode rejection ratio*, CMRR or CMR), will determine the effectiveness of this configuration in suppressing ground-loop interference. At higher frequencies, the cable capacitances will act with the unequal resistances in the A and B input circuits to form unequal low-pass filters, so that $e_A$ will no longer be equal to $e_B$ and there will be a spurious differential (A–B) input to the preamplifier. Though cable resistances and capacitances are shown for convenience as lumped parameters, it should be remembered that in fact they are distributed. As shown in Figure 11.9, high-frequency unbalance

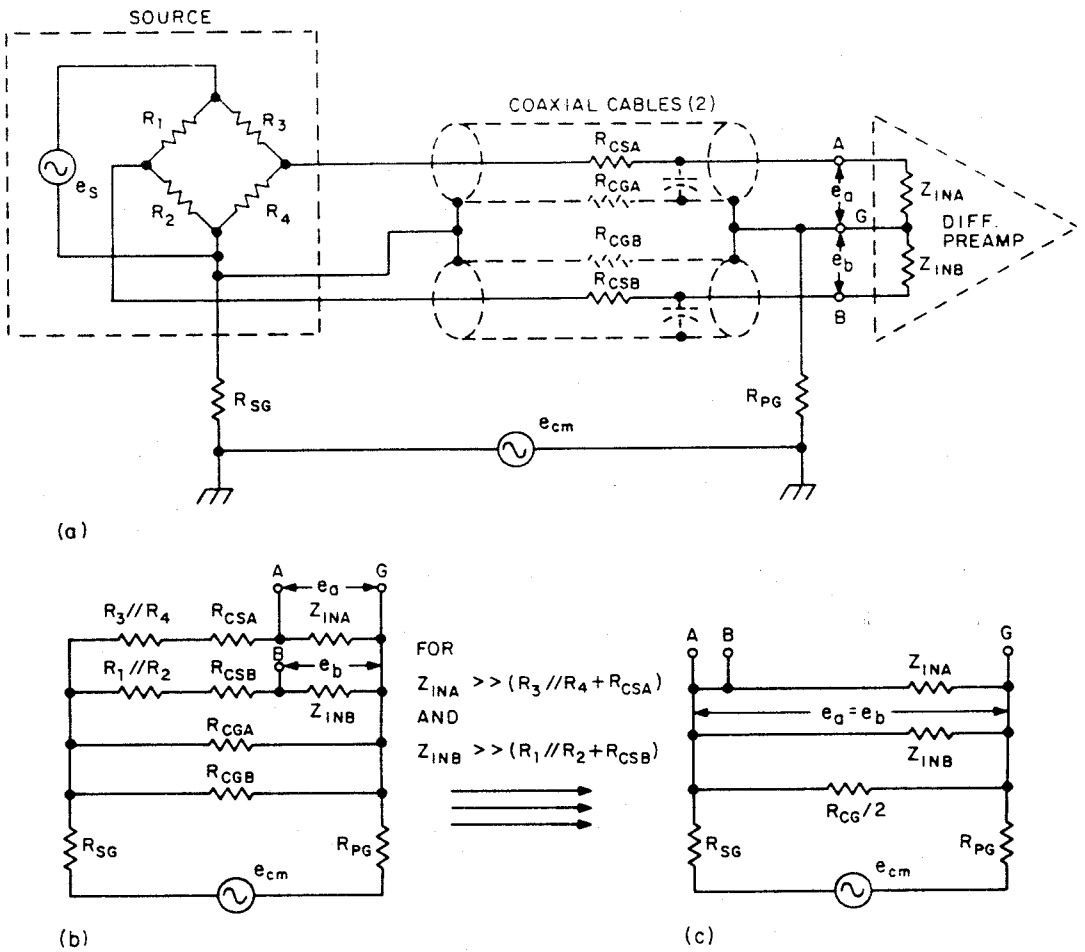



Figure 11.9 Differential preamplifier used with balanced source: (a) physical occurrence; (b) schematic equivalent circuit; (c) reduced equivalent circuit

problems can be avoided by using a balanced source. Specific comment on connecting stages together is to be found in (Morrison, 1977).

To end this section, here are a number of miscellaneous recommendations regarding good wiring and grounding practices.

(a) Keep cable lengths short; for differential connections, keep them equal and following the same route.

(b) Interference can be coupled into the ground (shield) or outer conductor of a coaxial cable. Consider coiling the cable to form an RF choke to suppress high-frequency interference of this kind, use a transformer, or use a balun (which allows d.c. continuity).

(c) Remember that a loop of wire acts as an antenna; reduce the area of such loops as much as possible.

(d) Separate low-level signals/cables from noisy ones. Where such cables must cross, cross them at right angles and with maximum separation.

(e) For non-coaxial connections use shielded twisted wire-pairs.
(f) Consider placing low-noise instruments in a shielded (screened) room when they are used with high-energy RF sources, such as pulsed lasers.
(g) Keep analog and digital grounds separate.

## 11.6 BANDWIDTH REDUCTION OF BASEBAND (d.c.) SIGNALS

The term *d.c. signal* is often used (and will be in this chapter) to mean a signal which has a frequency spectrum that includes zero frequency (d.c.). Technically, of course, a d.c. voltage or current is unvarying and, therefore, cannot carry information (other than that it exists). Such signals are also termed *baseband* signals, particularly when they are to be used to modulate a carrier frequency. The simplest way to improve the SNR for such signals is to use a low-pass filter to reduce the noise bandwidth to the point where any further reduction would also change the signal to an unacceptable extent.

Figure 11.10 shows a typical source and preamplifier system for such a pseudo-d.c. signal. We will use this circuit to show how the output SNR may by estimated and also how the SNR may be improved by reducing the noise bandwidth.

In this example, it is assumed that the photomultiplier tube (PMT) anode current consists of both a 5 Hz signal component ($i_s$ = 1 nA r.m.s.) and a d.c. component ($I_{d.c.}$ = 5 nA); typically, such d.c. currents are due to stray light and dark/leakage currents. The adjustable direct-current generator ($I_{zs}$) is used to null (zero offset) the d.c. component of the PMT current; that is, $I_{zs}$ is made equal and opposite to $I_{d.c.}$. This kind of zero suppression is often called *background subtraction.* Notice that $I_{zs}$ must be readjusted manually each time the background changes.

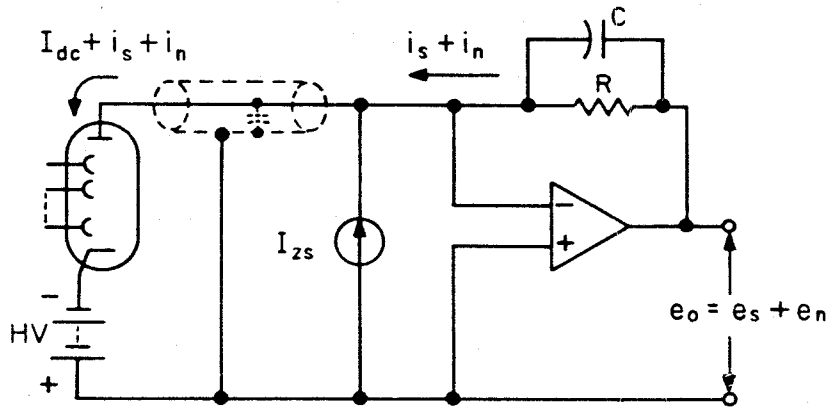



Figure 11.10 d.c. measurement example

The coaxial cable connecting the PMT and preamplifier has capacitance. Notice that the virtual-ground input of this preamplifier configuration offers the following advantages in addition to those discussed in Section 11.4:

(a) With zero volts across it, the cable capacitance cannot be charged and the cable will, therefore, be less microphonic than otherwise.
(b) Since the PMT anode voltage is clamped at zero volts, the anode-to-last-dynode voltage is also held constant regardless of $i_s$ (assuming that the dynode voltage remains fixed). Signal currents will, therefore, not change the PMT gain.

There are five, uncorrelated sources of noise in this circuit. These are:

(a) The d.c. component of the PMT current ($I_{d.c.}$) is produced by integrating anode pulses each of charge $Q = Ae$ where $A$ is the mean PMT gain. The interval in time between successive pulses arriving at the anode is random and governed by *Poisson* statistics (see Section 11.10.2). Assuming no additional dynode noise in the PMT, then the r.m.s. value of the PMT *shot noise* current spectral density, $i_{n1}$, is given by

$$i_{n1} = \sqrt{(2AeI_{d.c.})} = \sqrt{\text{A}} \times \sqrt{(2eI_{d.c.})}$$

For $A = 10^6$ (say), $R = 10^7\ \Omega$, $e \simeq 1.6 \times 10^{-19}$ C and $I_{d.c.} = 5$ nA, the resulting output noise voltage density, $e_{n1}$, is given by

$$e_{n1} = Ri_{n1} = 10^7 \times 10^3 \times \sqrt{2eI_{dc}} \simeq 10^{10} \times 4 \times 10^{-14}$$
$$= 4 \times 10^{-4} = 400\,\mu\text{V}/\sqrt{\text{Hz}}$$

(b) For purposes of this example, we can assume that the zero-suppress current, $I_{zs}$, is obtained from a transistor current source circuit so that it has a shot noise current spectral density, $i_{n2}$, given by

$$i_{n2} = \sqrt{2eI_{zs}} = \sqrt{(2eI_{d.c.})}(= i_{n1}/10^3) = 4 \times 10^{-14}\text{A}/\sqrt{Hz}$$

Note that though $I_{d.c.} = I_{zs}$, the shot noise component from the PMT is much larger than that from the transistor current source. The resulting output noise voltage spectral density, $e_{n2}$, is given by

$$e_{n2} = Ri_{n2} \simeq 10^7 \times 4 \times 10^{-14} = 4 \times 10^{-7} = 400\,\text{nV}/\sqrt{\text{Hz}}$$

(c) The feedback resistor, $R$, will generate (at $T = 290$ K) a Johnson noise output voltage density, $e_{n3}$, given by

$$e_{n3} = \sqrt{(4kTR)} \simeq 4 \times 10^{-7} = 400\,\text{nV}/\sqrt{\text{Hz}}$$

(d) At 5 Hz, a typical value for the spot noise voltage density of the amplifier's internal noise voltage generator is 30 nV/$\sqrt{\text{Hz}}$. This amplifier voltage noise will experience unity gain (see Figure 11.3b) since $Z_{in}$ (the PMT

current source) is very high. The output noise voltage density, $e_{n4}$, due to this noise source is therefore given by

$$e_{n4} = 30\ \text{nV}/\sqrt{\text{Hz}}$$

(e) At 5 Hz, a typical value for the spot noise current density, $i_{n5}$, of the amplifier internal noise current generator is 5 fA/$\sqrt{\text{Hz}}$. The resulting contribution, $e_{n5}$, to the amplifier output noise is given by

$$e_{n5} = Ri_{n5} = 10^7 \times 5 \times 10^{-15} = 5 \times 10^{-8} = 50\ \text{nV}/\sqrt{\text{Hz}}$$

The total output noise voltage spectral density, $e_n$, is given by

$$e_n = \sqrt{(e_{n1}^2 + e_{n2}^2 + e_{n3}^2 + e_{n4}^2 + e_{n5}^2)}$$

Since $e_{n1}^2 \gg e_{n2}^2, e_{n3}^2, e_{n4}^2$, and $e_{n5}^2$, then

$$e_n \simeq e_{n1}$$

and the system is said to be *detector limited* or *shot noise limited.* An *electrometer*, an instrument characterized by extremely low leakage currents, is often used as a low-noise amplifier in d.c. measurements of this kind.

In Figure 11.10 the parallel resistor and capacitor in the feedback loop cause the circuit to act as low-pass filter of time constant $RC$ seconds, so that the $-3$ dB cutoff frequency is given by $1/2\pi RC$ and $B_n = 1/4RC$ (see Section 11.2).

If no discrete capacitor is connected across $R$, the typical stray capacitance will (say) be about $C = 2.5$ pF, so that $RC = 10^7 \times 2.5 \times 10^{-12} = 25\ \mu$s and $B_n = 10^4$ Hz. The output noise voltage ($E_n$) will, therefore, be

$$E_n = e_n\sqrt{B_n} = 4 \times 10^{-4} \times \sqrt{10^4} = 40\ \text{mV}$$

The output signal

$$e_s = i_s R = 10^{-9} \times 10^7\ \text{V} = 10\ \text{mV}$$

Therefore,

$$\text{SNR} = \frac{S}{N} = \frac{10}{40} = \frac{1}{4}$$

The capacitance can be increased to 2.5 nF by adding discrete capacitors so that the noise bandwidth becomes $B_n = 10$ Hz. The $-3$ dB corner frequency will now be at 6.4 Hz (i.e. $10 \times 2/\pi$) so that the signal (frequency is 5 Hz) is not significantly attenuated. The output noise voltage ($E_n$) is now reduced to

$$E_n = e_n\sqrt{B_n} = 4 \times 10^{-4} \times \sqrt{(10)} \simeq 1.26\ \text{mV}$$

and

$$\text{SNR} = \frac{S}{N} = \frac{10}{1.26} \simeq \frac{8}{1}$$

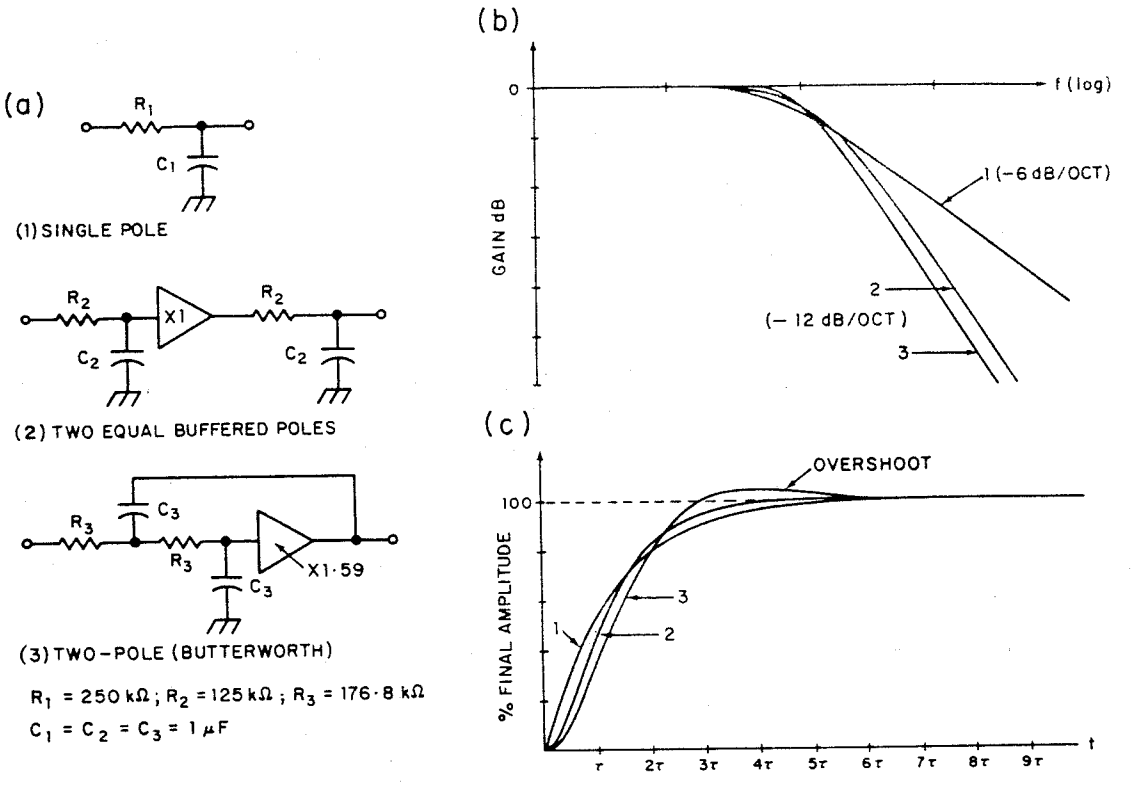



Figure 11.11 Low-pass filter characteristics: (a) filter types (all have same ENBW); (b) frequency responses; (c) time responses to voltage step input

so that (see equations (11.9) and (11.10))

$$\text{SNIR} = \frac{S_o/N_o}{S_i/N_i} = \sqrt{\frac{B_{ni}}{B_{no}}} = \sqrt{\frac{10000}{10}} \simeq \frac{32}{1}$$

The roll-off rate of a low-pass filter may be increased by adding more *RC* sections (see Figure 11.11a). Care should be taken in using some multipole filter configurations (Chebyshev or Butterworth, for example), since many such filters have undesirable overshoot characteristics (see Figure 11.11c). Notice that the term *time constant* ($\tau$) is meaningful only in connection with a single *RC* filter section and, even then, does not adequately convey a sense of the response time of the filter. With a voltage-step input, for example, such a single *RC* section requires about five time-constant intervals for its output to rise to within 1% of its final value.

## 11.7 AMPLITUDE-MODULATED SIGNALS; THE LOCK-IN AMPLIFIER

Most measurement systems are troubled by $1/f$ noise. By amplitude modulating the measurand (quantity to be measured) at some reference or carrier frequency, $f_r$, the output noise can often be reduced and d.c. drift problems avoided (see

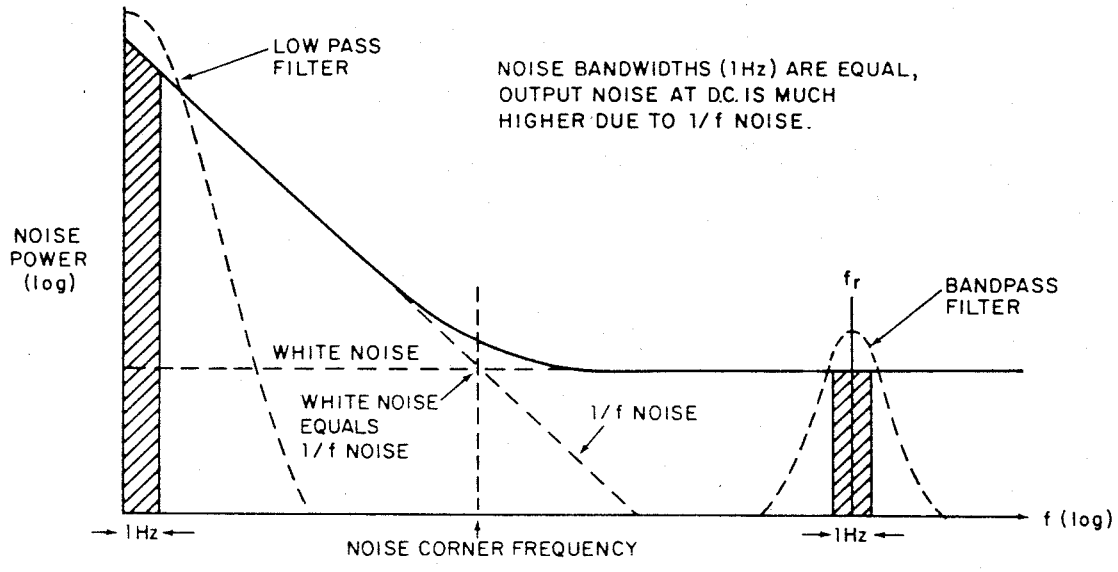



Figure 11.12 Amplitude modulation to avoid $1/f$ noise

Figure 11.12). In optical systems, for example, rotating or vibrating mechanical chopper blades are often used to periodically block a light beam and thereby square-wave modulate the signal amplitude—even though, in most cases, such chopping means losing half of the light (signal). Measuring instruments that respond only to the modulation provide *automatic background subtraction*; as with d.c. systems, however, the noise component of the background remains. Such modulation also allows the use of transformers to noise-match preamplifiers to low-resistance sources.

As with baseband signals and low-pass filtering, the SNR of a noisy amplitude-modulated signal can be improved by bandwidth reduction—in this case a *band-pass* filter is commonly used. In most applications, carrier frequencies are chosen from the 100 Hz–10 kHz range, where preamplifier and environmental noise is lowest; care should also be taken to avoid frequencies occupied by harmonics of the power-line frequency. A second-order band-pass filter (see Figure 11.13) is specified by its resonant or centre frequency, $f_r$, and its selectivity, $Q$ (quality factor). For a given value of $f_r$, the higher the $Q$, the narrower the filter width.

The $-3$ dB frequencies are at $f_r \pm f_c$ and signal bandwidth ($B_s$) is defined by

$$B_s = 2f_c = f_r/Q \tag{11.14}$$

For a second-order band-pass, the signal bandwidth and the equivalent noise bandwidth ($B_n$) are related by

$$B_n = \tfrac{1}{2}\pi B_s$$

so that,

$$B_n = \pi f_r/2Q \tag{11.15}$$

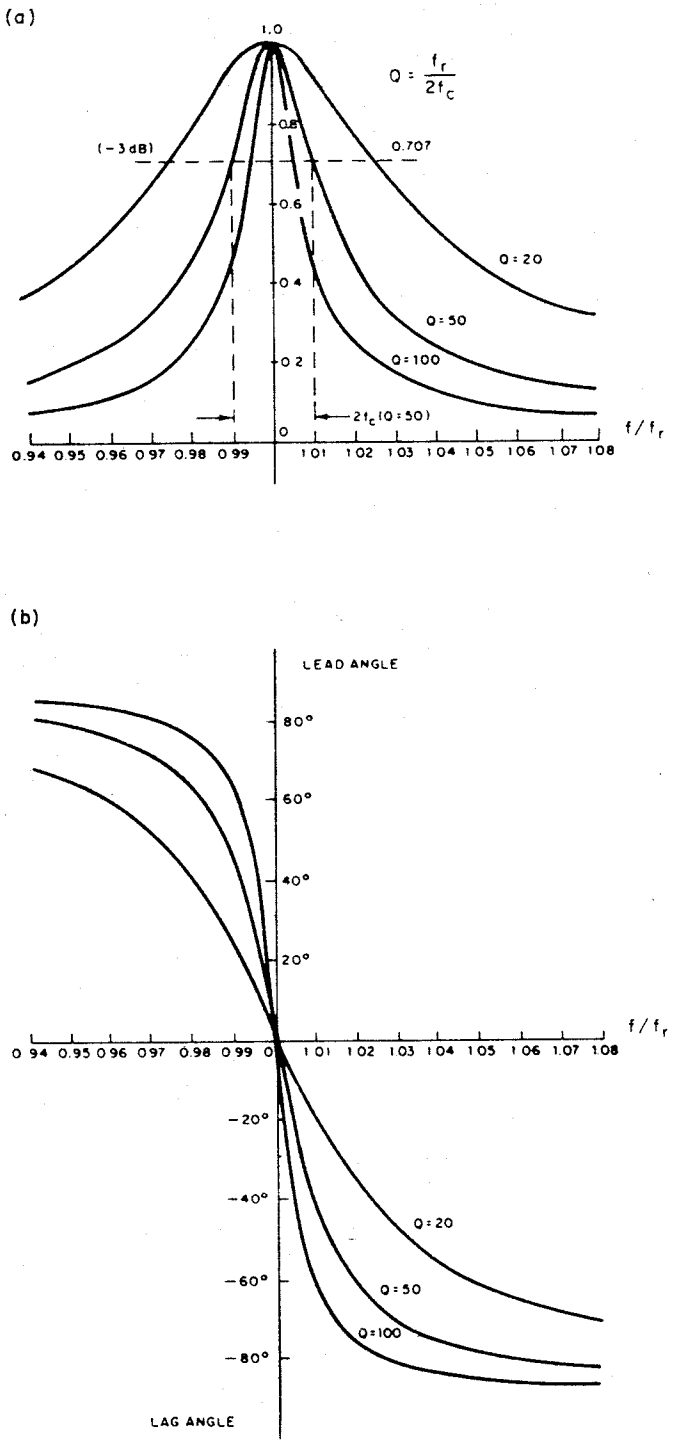



Figure 11.13 Characteristics of a second-order bandpass filter: (a) amplitude; (b) phase

The band-pass filter has associated with it an effective time constant, $\tau$, where (as with the low-pass filter case discussed in Section 11.2)

$$f_c = \frac{1}{2\pi\tau} \tag{11.16}$$

so that,

$$B_n = \tfrac{1}{2}\pi B_s = \tfrac{1}{2}\pi 2 f_c = \tfrac{1}{2}\pi(2/2\pi\tau) = 1/2\tau \tag{11.17}$$

Also, from equations (11.14) and (11.16)

$$\tau = 1/2\pi f_c = Q/\pi f_r \tag{11.18}$$

From equation (11.15), we can see that the higher the $Q$, the smaller the noise bandwidth and, therefore, for white noise or other broad-band noise interference, the smaller the noise and the better the SNR. With a band-pass filter implemented by active $RC$ (or $LC$) circuitry, frequency-stability problems limit the maximum practicable value of $Q$ to about 100.

The *lock-in amplifier* (EG & G, 1) is in part, a band-pass filter–amplifier that overcomes the $Q$ limitations of conventional circuits; noise bandwidths of less than 0.001 Hz and $Q$ values of $10^8$ or greater are easy to implement. The lock-in amplifier can also provide amplification of more than $10^9$ (180 dB). The term *lock-in* comes from the fact that the instrument locks in to the frequency ($f_r$) of a reference signal. With an external reference signal, a lock-in acts as a tracking band-pass filter and detector with a centre frequency equal to that of the reference ($f_r$); it will automatically track changes in $f_r$ and can be used in a frequency-scanning mode if desired. Commercial instruments are available to cover a frequency range of about 0.1 Hz–50 MHz.

Though not all lock-in amplifiers use a phase-locked loop in their reference channel, most single-phase lock-ins may be represented by the simplified block diagram shown in Figure 11.14. The reference input waveform to the

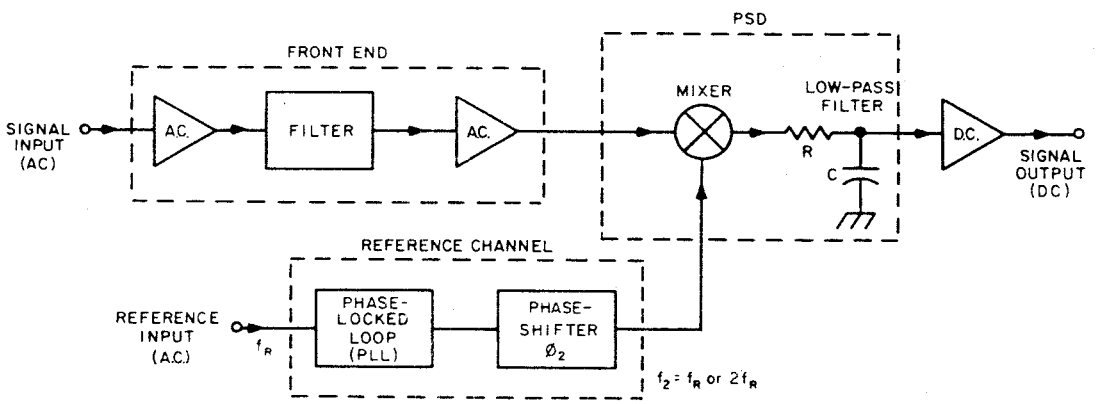



Figure 11.14 Basic lock-in amplifier (simplified) (reproduced by permission of EG&G Princeton Applied Research Corporation

lock-in may be of almost any waveshape and its zero crossings are used to define zero phase ($\phi_r = 0$). The output of the phase-locked loop circuit is a precise square-wave, locked in phase to the reference input, and at a frequency $f_2$. Normally, $f_2 = f_r$ (the reference frequency); most lock-ins also provide a second-harmonic mode where $f_2 = 2f_r$ and this mode is often used for derivative (signal rate of change) measurements.

All lock-ins use a *phase-sensitive detector* (PSD) circuit and all PSD circuits consist of nothing more than a *mixer* followed by a low-pass filter. The output of a mixer ($e_3$) is the product of its signal ($e_1$) and gating ($e_2$) inputs, that is, $e_3 = e_1 e_2$, and the phase difference between these two inputs can be precisely adjusted by the phase-shifter circuit in the reference channel. For use in lock-in amplifiers, mixer circuits must be capable of withstanding large amounts of noise (i.e. asynchronous signals, $f_1 \neq f_2$) without overloading. The term *dynamic reserve* is used to specify such noise overload performance (see Figure 11.15). Dynamic reserve is defined as the ratio of the overload level (peak value of an asynchronous signal that will just cause significant non-linearity), to the peak value of a full-scale synchronous signal. Dynamic reserve is often confused with *dynamic range*, which is the ratio of the overload level to the minimum detectable signal level.

The d.c. drift of both the mixer and d.c. amplifier may limit the minimum detectable signal and the gain of the d.c. amplifier should, therefore, be minimized to provide optimum output stability; a.c. gain should be used to provide most of the overall instrument gain required. Such a gain distribution is practicable and desirable for use with clean signals. With noisy signals, however, the a.c. gain must be reduced to provide increased dynamic reserve, and the d.c. gain increased proportionately. Most high-performance instruments

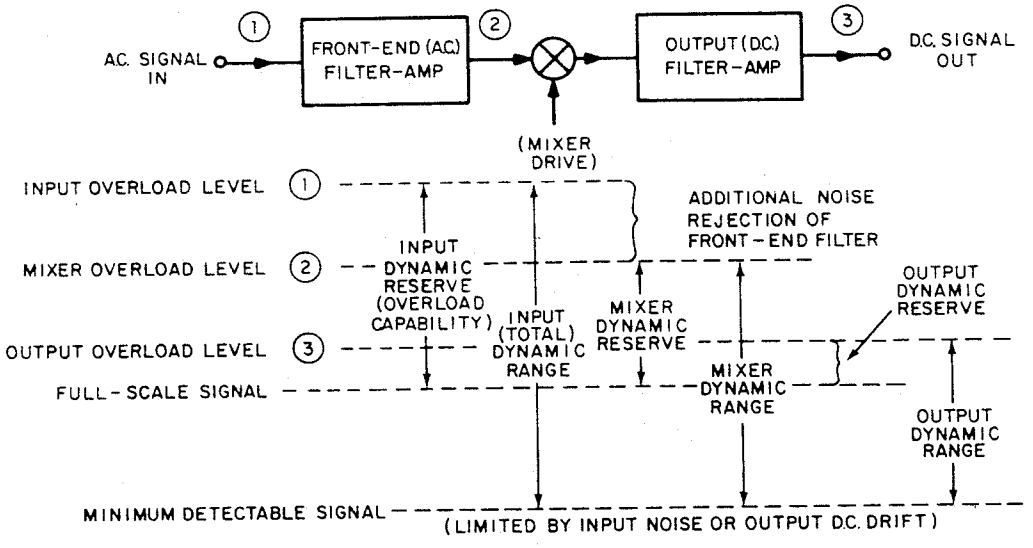



Figure 11.15 Dynamic range and dynamic reserve of a lock-in amplifier

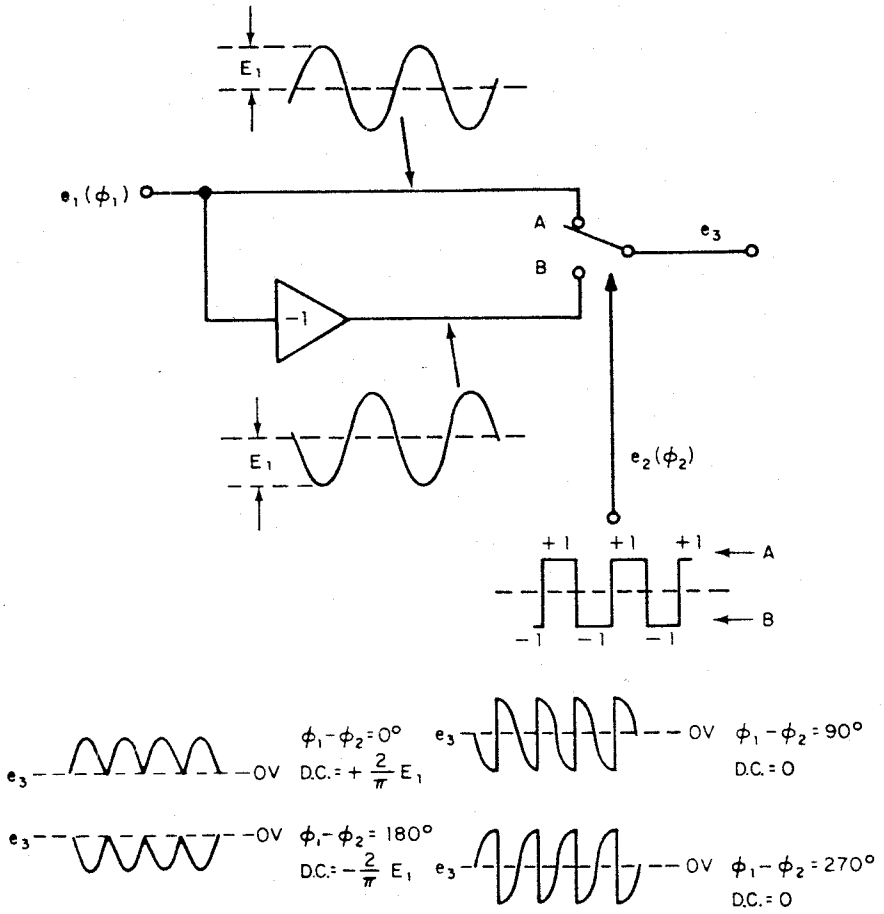



Figure 11.16 Switching mixer waveforms (reproduced by permission of EG&G Princeton Applied Research Corporation)

provide the controls to allow such a trade-off between dynamic reserve and output stability.

Due to non-linearity and other problems, a linear multiplier type of mixer cannot provide the dynamic range required in a (commercial) lock-in amplifier. Consequently, mixers are invariably of the switching type, shown in Figure 11.16. The switch shown in this figure will be in position A during positive half-cycles of the square-wave drive waveform and in position B during negative half-cycles. When the signal and drive waveforms have a common frequency component, as shown, the mixer acts as a synchronous rectifier and produces a phase-sensitive d.c. output. Outputs are shown for four different phase relationships. Notice that the mixer d.c. output can be adjusted from zero to $\pm(2/\pi)E$, by varying the phase-difference $(\phi_1 - \phi_2)$. The square-wave drive has an effective amplitude of $\pm 1$ and contains all odd harmonics of the fundamental frequency of the square-wave. The output of a mixer, therefore, is composed of a large number of frequencies (see Figure 11.17). Thus, $f_1 + f_2, f_1 + 3f_2, f_1 + 5f_2, \ldots$, are sum frequencies; $f_1 - f_2, f_1 - 3f_2, f_1 - 5f_2, \ldots$,

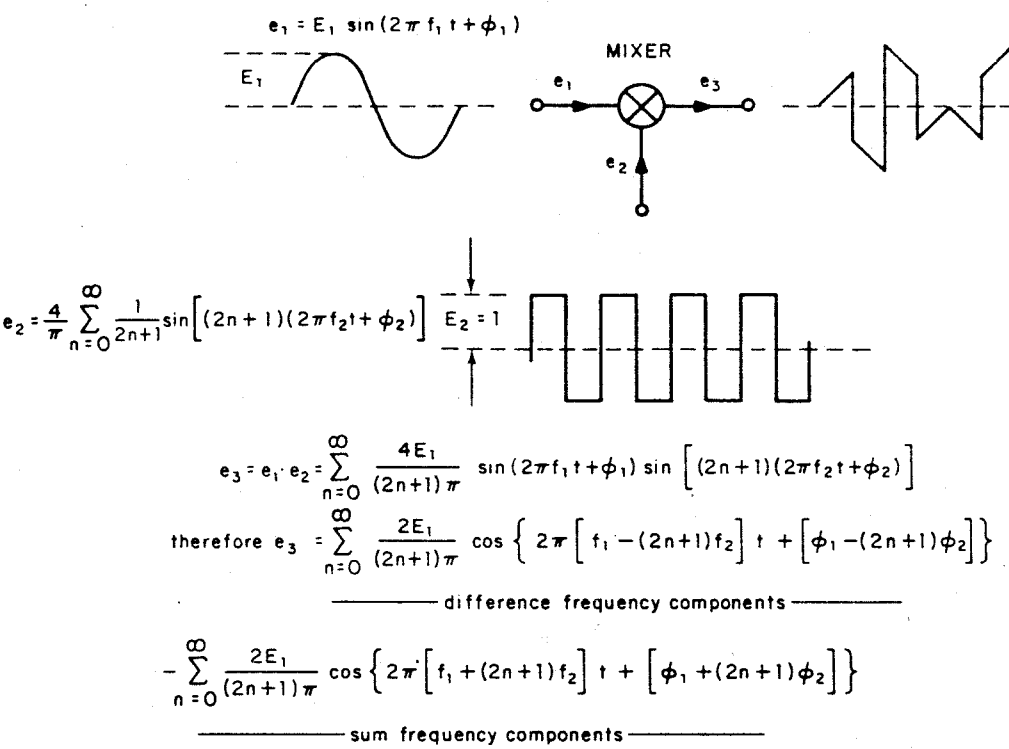



Figure 11.17 Switching mixer operation (reproduced by permission of EG&G Princeton Applied Research Corporation)

are difference frequencies. Note that when $f_1 = (2n + 1)f_2$, one of the difference-frequency components of the mixer output will be at zero frequency, or d.c.. A mixer will, therefore, produce a phase-sensitive d.c. output when $f_1 = (2n + 1)f_2$ and for $n > 0$, these outputs have magnitudes that are inversely proportional to their harmonic number ($n$) and are known as the *harmonic responses* of the mixer. Lock-ins respond to the *average full-wave rectified* value of the input signal but are usually calibrated in r.m.s.—a sinusoidal input or sinusoidal front-end response is assumed.

The PSD input ($e_1$) need not be sinusoidal. If $e_1$ were a synchronous square-wave signal, for example, such as that resulting from chopped-light experiments, then $e_1$ would contain a large number of synchronous components each of which would give rise to an output d.c. signal from the PSD.

In a perfect mixer, only synchronous inputs can cause a d.c. output. In practice, due to non-linearities of the mixer switching elements, a mixer can produce a d.c. output with high-level noise inputs; even with no (zero) input, capacitive feedthrough can cause a d.c. output. Such spurious d.c. outputs are normally negligible in amplitude. However, at higher frequencies (above 10 kHz typically), the magnitude of such a d.c. offset and its associated drift may become significant.

As we saw previously in Figure 11.17, the output of a mixer contains a large number of sinusoidal sum and difference frequency components. (The number

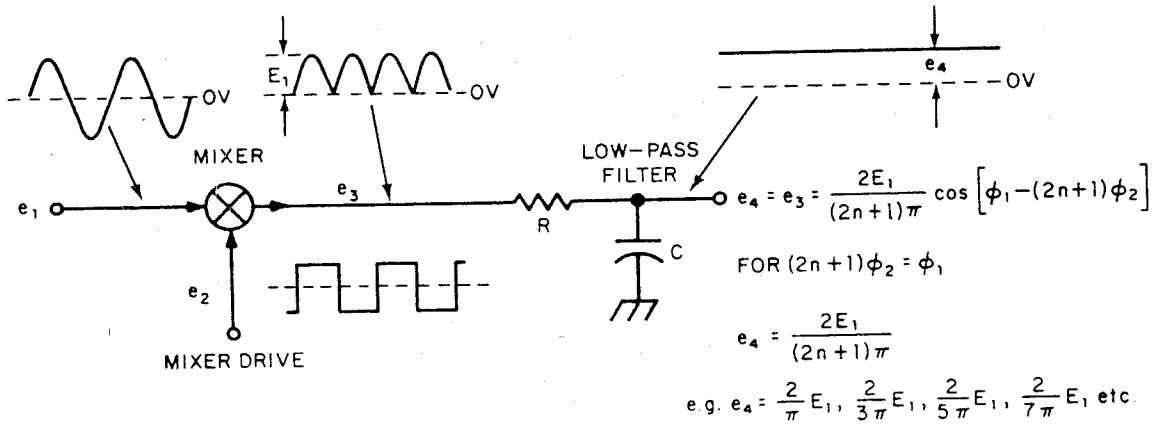



Figure 11.18 PSD operation with synchronous signal (reproduced by permission of EG&G Princeton Applied Research Corporation)

is large rather than infinite, since the squareness of the drive signal is not perfect, and the drive signal does not contain all higher odd harmonics). The effect of the low-pass filter (see Figure 11.18) is to remove all components of the mixer output which have frequencies beyond the filter cut-off. When the filter time constant is set normally (so that the filter cut-off frequency ($f_c$) is appreciably less than the fundamental frequency ($f_2$) of the mixer square-wave drive), the output of a PSD will contain only those difference frequency components having frequencies within (approximately) the equivalent noise bandwidth of the filter.

Suppose, as shown in Figure 11.19, that the PSD input ($e_1$) is asynchronous (noise) of frequency $f_1 = f_2 + \Delta f$. The resulting mixer sum and difference frequencies (ignoring harmonics for simplicity) will, therefore, be $2f_2 + \Delta f$ and $\Delta f$ respectively. Only the $\Delta f$ component may be low enough in frequency to pass through the low-pass filter and appear as output noise. Suppose we change the frequency of this input noise to $f_1 = f_2 - \Delta f$. The resulting sum and difference frequencies will respectively be $2f_2 - \Delta f$ and $-\Delta f (= \Delta f)$. Again, only the $\Delta f$ component can appear as output noise and the low-pass filter 'cannot tell' whether its $\Delta f$ input resulted from an $f_2 + \Delta f$ input to the mixer or an

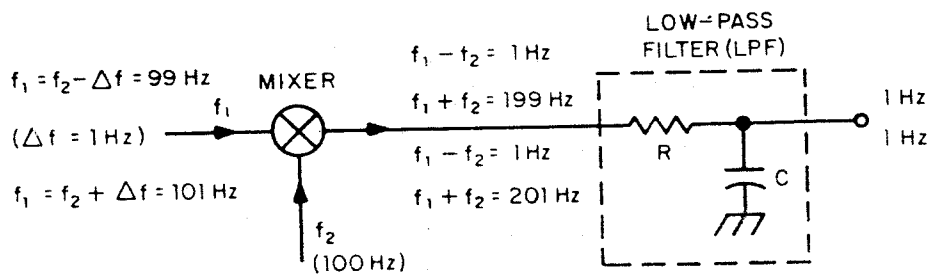



Figure 11.19 PSD operation with asynchronous (noisy) signal (reproduced by permission of EG&G Princeton Applied Research Corporation)

$f_2 - \Delta f$ input. In addition to its rectifying and phase-sensitive properties, therefore, the PSD filters noise as though it consisted of *band-pass* responses centered on all odd harmonics of $f_2$ (see Figure 11.20). Notice that each effective band-pass response consists of the output low-pass filter response and its mirror image; their centre-frequencies automatically track changes in the PSD drive frequency $f_2$.

Each band-pass response has an equivalent noise bandwidth determined by that of the low-pass filter. If the PSD input consists of white noise, the effect of the harmonic responses ($2n + 1 = 3, 5, 7, \ldots$) is to increase the PSD output noise by 11%. For square-wave signal inputs, the additional output noise (11%) caused by the harmonic responses is more than compensated for by the increase in signal (23%). If the PSD is used to measure a sinusoidal signal accompanied by white noise, a separate low-pass or band-pass filter, centred on $f_2$, may be used in front of the PSD to remove the harmonic responses and thus the additional 11% noise. The improvement in output SNR effected by the use of such front-end filtering on white noise is normally insignificant. With discrete frequency noise, however, front-end filters can be extremely helpful. By reducing the input noise before it reaches the PSD, the overload

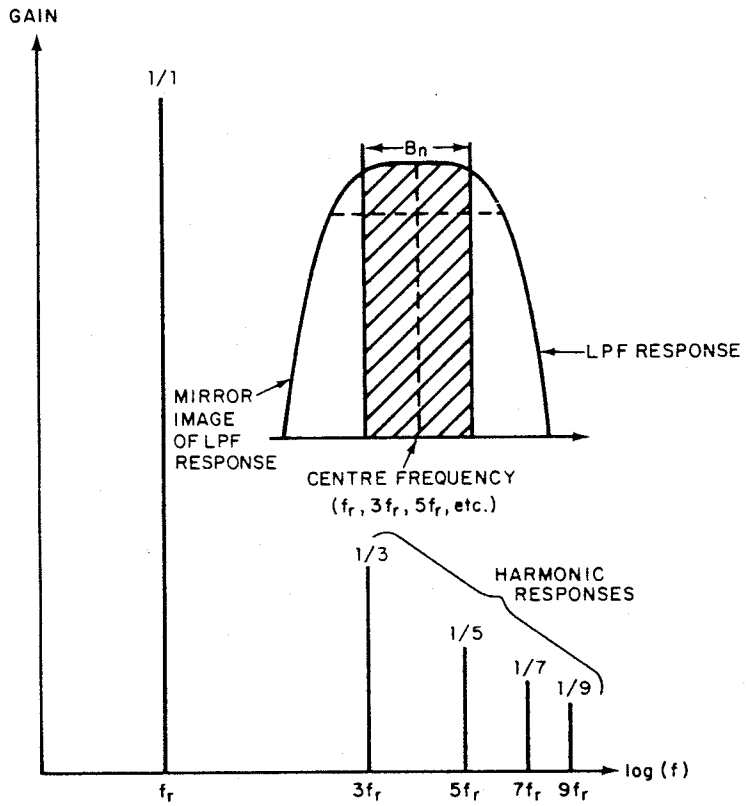



Figure 11.20 Frequency response of a PSD (reproduced by permission of EG&G Princeton Applied Research Corporation)

capability of the lock-in may be improved significantly beyond the dynamic reserve of the mixer and this additional dynamic range can also be used to provide increased output stability if desired. For this reason some lock-ins provide adjustable low-pass, high-pass, band-pass and band-reject (notch) filters, in addition to a flat frequency response (broad-band) mode. Applications and explanation of operation of PSD systems are also reviewed in Blair and Sydenham (1975).

Heterodyning front-ends are also available. With this approach, a fixed-frequency band-pass amplifier is used to protect the PSD and increase the overload capability of the lock-in. In order to use such a fixed-frequency filter, another mixer is used to heterodyne the input signal frequency up to the centre frequency of the filter; two such heterodyning schemes are shown in Figure 11.21. The advantage of this approach is that the instrument offers sinusoidal response (no harmonic responses) and overload performance approaching that of a tunable band-pass instrument, without the need for manual tuning. Because of the phase-shift characteristic (see Figure 11.13) of their front-end band-pass filters, however, manually tuned or heterodyning instruments cannot provide the phase stability of a broad-band or flat lock-in.

Figure 11.22 shows a simplified block diagram of a *two-phase* or *vector* lock-in amplifier. An additional quadrature ($Q$) output channel has been added, consisting of a mixer, low-pass filter, and d.c. amplifier. The reference channel provides quadrature gating inputs to the $I$ (in-phase) and $Q$ mixers. The *orthogonality* of the two mixer drives—that is, the accuracy of the 90° phase difference between them—is extremely important when measuring a small $I$ signal in the presence of a large $Q$ signal (or vice versa). Similarly, in a single-phase lock-in, the accuracy with which a 90° phase shift can be switched into or out of circuit (using the phase-quadrant switch), is equally important.

Most two-phase lock-ins provide a vector/phase circuit (usually as an option), which computes the vector magnitude $M$, where

$$M = \sqrt{(I^2 + Q^2)} = \sqrt{[(A\cos\phi)^2 + (A\sin\phi)^2]} = A$$

where $A$ is the signal amplitude, and

$$\phi = \tan^{-1}\left(\frac{Q}{I}\right) = \tan^{-1}\left(\frac{A\sin(\phi_s + \phi_r)}{A\cos(\phi_s + \phi_r)}\right) = \phi_s + \phi_r$$

where $\phi_s$ is the signal phase shift relative to the reference signal phase, and $\phi_r$ is the phase offset set by the phase-shift controls. Two-phase lock-ins can, therefore, display their output signal in rectangular or polar form, with the phase controls ($\phi_r$) allowing continuous vector rotation. Notice that asynchronous signals ($f_s \neq f_r$) with beat frequencies within the low-pass filter response will provide d.c. outputs and the instrument, therefore, acts as a *wave analyser*. Modern wave analysers are essentially vector lock-ins that are

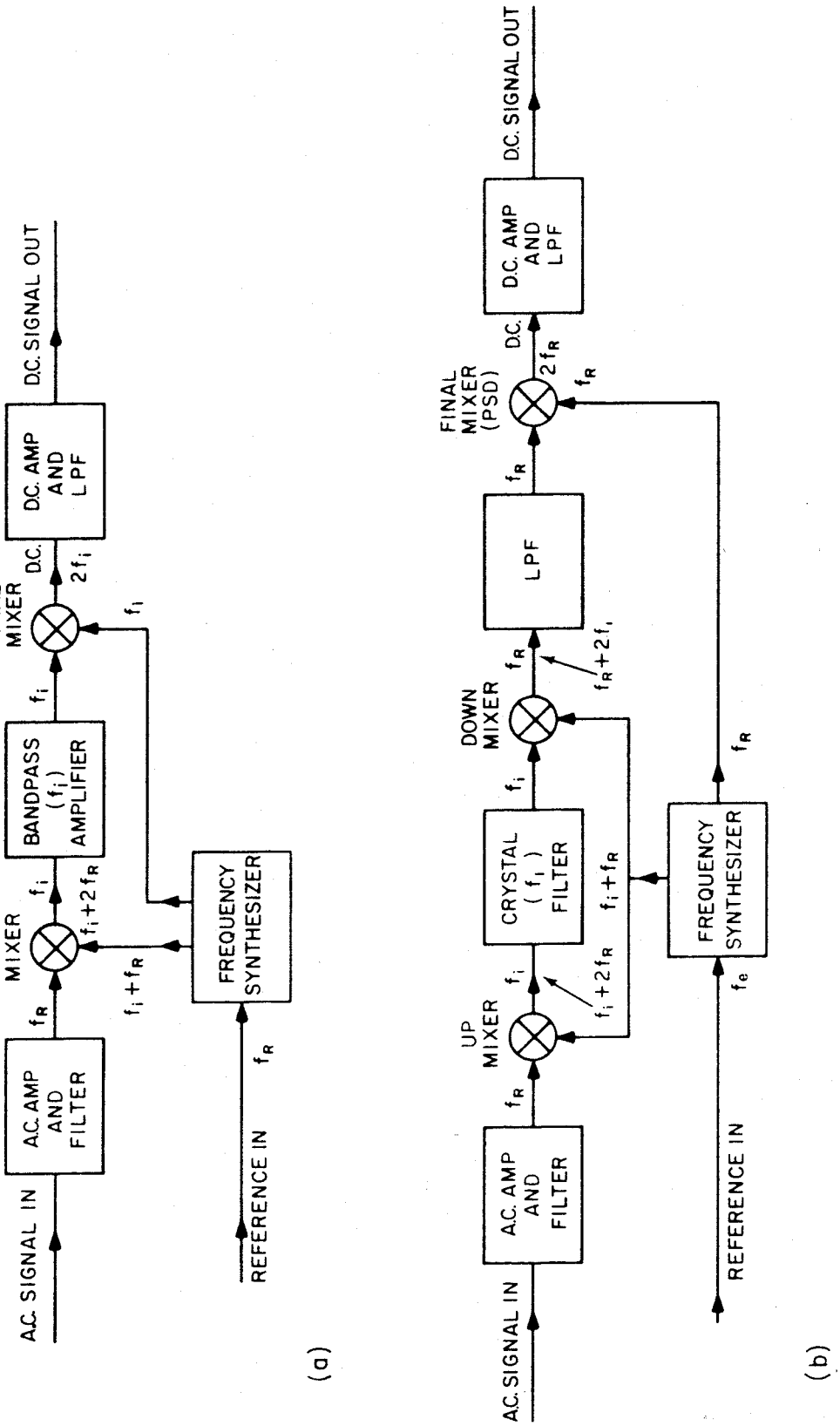



Figure 11.21 Heterodyning lock-in amplifiers: (a) single up-conversion; (b) double heterodyning

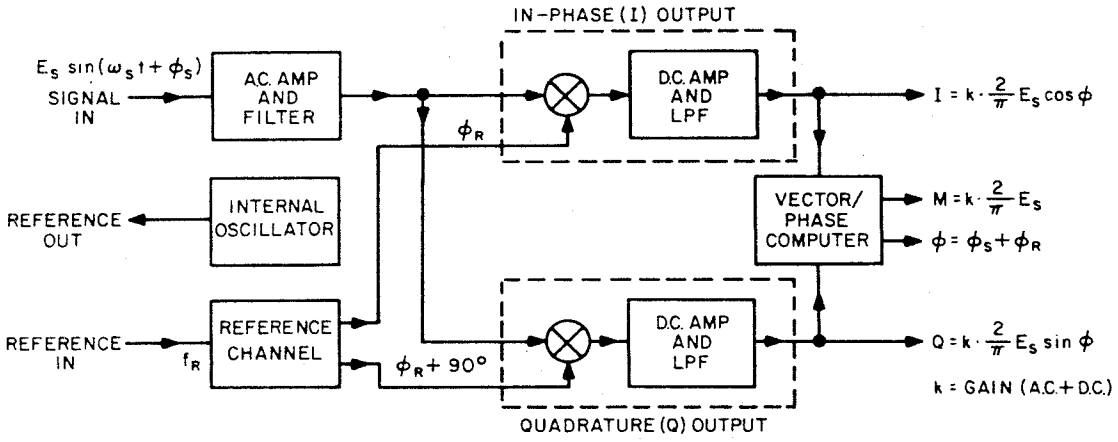



Figure 11.22 The two-phase (vector) lock-in amplifier

optimized for convenience in measuring frequency components of a signal rather than recovering a signal from noise.

Figure 11.23 shows a typical application for a two-phase lock-in. In such a.c. bridge applications, the phase shift ($\phi_r$) can be set to zero, so that the in-phase ($I$) signal responds only to the bridge resistance and the quadrature output ($Q$) to the bridge capacitance. The bridge can then be balanced very simply by separately nulling $R_s$ and $C_s$.

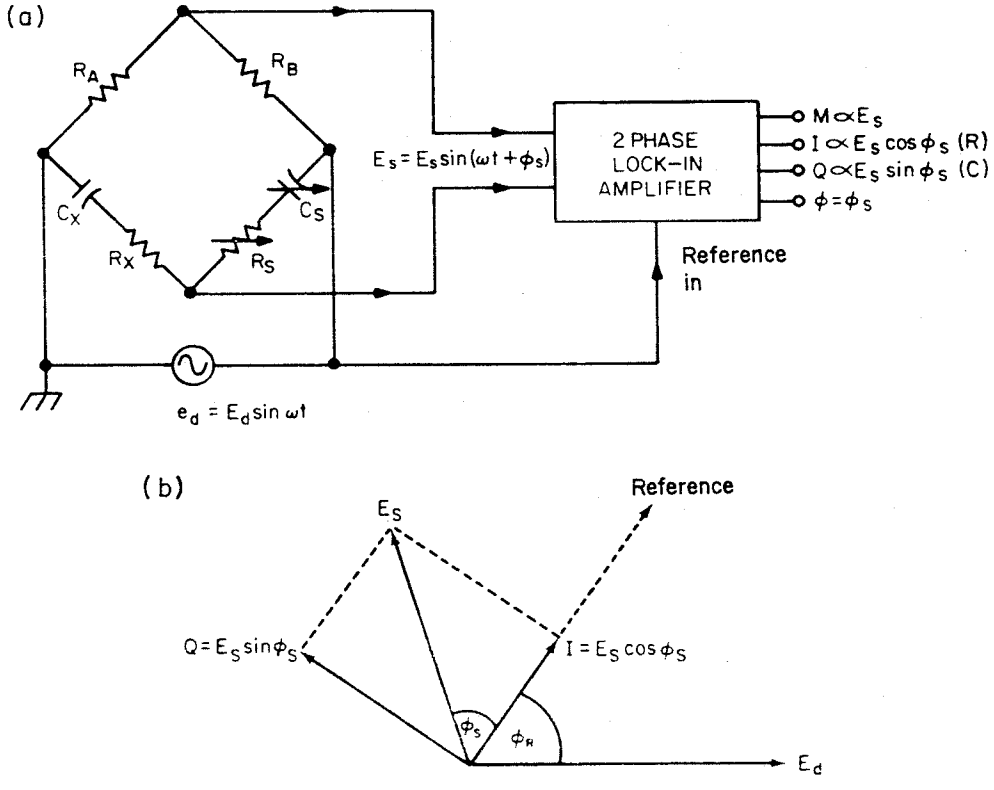



Figure 11.23 Alternating current bridge measurements with a two-phase lock-in amplifier: (a) system arrangement; (b) vector relationships

As in the case of d.c. measurements with preamplifiers or electrometers, the SNIR to be expected from a lock-in depends upon the input noise bandwidth ($B_{ni}$) to the lock-in, the noise bandwidth ($B_{no}$) of the lock-in, and the noise spectral characteristics. For random white noise and unity gain,

$$\text{SNIR} = \sqrt{(B_{ni}/B_{no})}$$

where for a $-6$ dB/octave rate of roll-off for the lock-in output low-pass filter, $B_{no} = 1/4RC$, or for a $-12$ dB/octave low-pass filter roll-off, $B_{no} = 1/8RC$.

## 11.8 SIGNAL AVERAGING

### 11.8.1 The Boxcar Averager

The *boxcar averager* (EG & G, 2) is a sampling instrument that is used to enhance repetitive signals. Also known as a *boxcar integrator* or *detector*, the boxcar takes only one sample during each signal occurrence or sweep, and requires a trigger signal at a fixed time interval prior to the beginning of each such sweep. The heart of any boxcar is the *gated integrator* circuit, shown in simplified form in Figure 11.24. This circuit is simply an $RC$ low-pass filter gated by switch $S_1$ (the sampling gate). As shown, the gated integrator has unity d.c. signal gain.

If the gate is opened (i.e. $S_1$ *closed*) every $T$ seconds for an aperture of $t_g$ seconds, then the duty factor $\gamma$ is given by $\gamma = t_g/T = t_g f$ where $f = 1/T$. When $e_i$ is a voltage step, $e_o$ will rise exponentially as shown in Figure 11.25, curve A. Notice that the *effective* time constant $\tau_{eff}$, is much longer than the real (ungated) time constant, $RC$, and is given by

$$\tau_{eff} = \frac{RC}{\gamma} = \frac{RC}{t_g f} \tag{11.19}$$

As we saw previously for the PSD of a lock-in amplifier, the noise bandwidth ($B_{no}$) of the gated integrator is simply that of the low pass filter, that is $B_{no} = 1/4RC$.

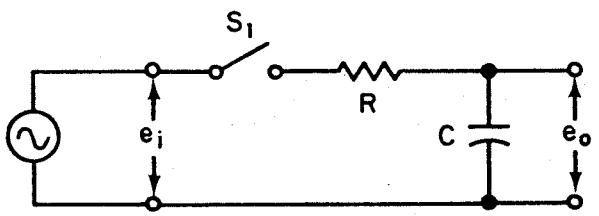



Figure 11.24 The gated integrator (simplified)

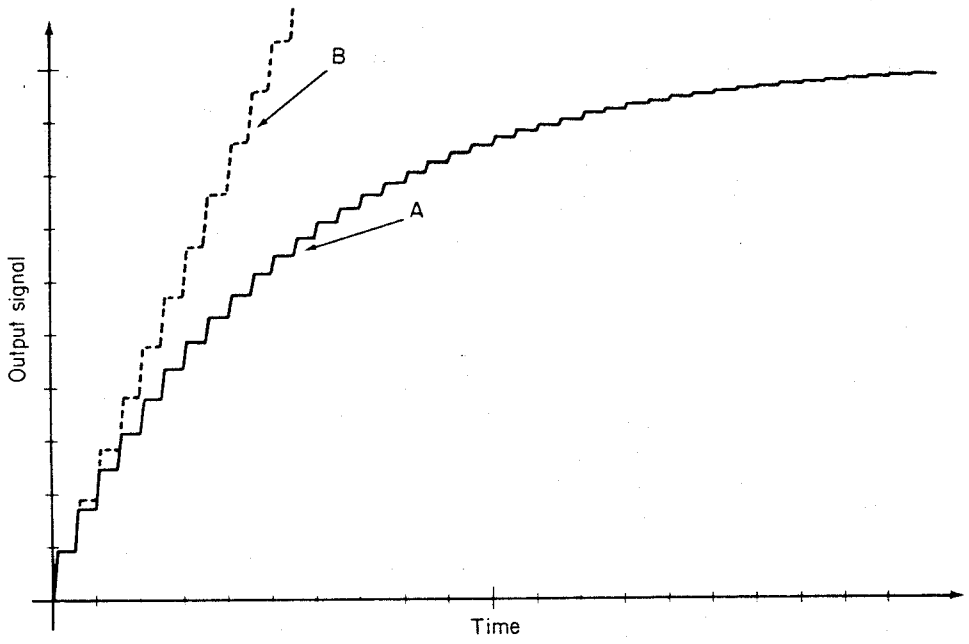



Figure 11.25 Gated integrator modes of operation: curve A, Exponential averaging $\tau = RC/(t_s f)$; curve B, Linear summation

For input white noise limited to a bandwidth $B_{ni}$ and unity gain, and where

$$B_{no} \ll B_{ni} \quad \text{and} \quad B_{no} \ll 1/t_g$$

then

$$\text{SNIR} = \sqrt{(B_{ni}/B_{no})} = \sqrt{(4RCB_{ni})} \tag{11.20}$$

The time resolution of a boxcar measurement will improve (shorten) as the gate duration ($t_g$) is decreased, until the point is reached where the input bandwidth ($B_{ni}$) limits the resolution. If we set

$$t_g = 1/(2B_{ni}) \quad \text{or} \quad B_{ni} = 1/(2t_g)$$

then we obtain the widely quoted formula

$$\text{SNIR} = \sqrt{(4RCB_{ni})} = \sqrt{(2RC/t_g)} \tag{11.21}$$

With pulsed signals, the ability of a boxcar to separate temporally the signal from (most of the) noise is usually of much greater significance than such theoretical white noise considerations.

In the *exponential averaging* or exponential weighting mode shown in Figure 11.24, the output signal from the gated integrator favours the most recent samples and provides a dc output that follows the input at a reasonable rate. In many ways this mode of boxcar operation resembles a lock-in amplifier; in fact, if two gated integrator channels are used, one to sample signal plus background, the other set to sample the background only, then by taking the difference

between the two outputs, we have essentially built a lock-in amplifier with adjustable duty cycle.

Figure 11.26 shows a simplified schematic of a complete boxcar averager. When switch $S_4$ is moved to the A position, the circuit behaves as a true gated integrator rather than as a gated low-pass filter. In this mode, all samples have equal weight and, for a step input, the output will rise in a linear staircase fashion as shown in Figure 11.25, curve B. In this *linear summation* mode, the desired number of signal samples ($m$) is selected; after $m$ triggers have occurred, switch $S_3$ is used to reset the integrator (discharge capacitor $C$). Since the signal samples will add linearly, and random noise samples will add vectorially, after $m$ samples of a constant amplitude signal ($S$) plus white noise ($N$), and after maximizing the gate width to suit the signal waveshape, the output SNR is given by:

$$\mathrm{SNR_{out}} = \frac{S_1 + S_2 + S_3 + \cdots + S_m}{\sqrt{(N_1^2 + N_2^2 + N_3^2 + \cdots + N_m^2)}} = \frac{mS}{\sqrt{(mN^2)}} = \frac{S}{N}\sqrt{m} = \mathrm{SNR_{in}}\sqrt{m}$$

so that

$$\mathrm{SNIR} = \frac{\mathrm{SNR_{out}}}{\mathrm{SNR_{in}}}\left(= \frac{\mathrm{SNR}\ (m\ \text{samples})}{\mathrm{SNR}\ (1\ \text{sample})}\right) = \sqrt{m} \qquad (11.22)$$

Notice that for this operating mode, it is easiest to think in terms of time averaging since the equivalent noise bandwidth of the gated integrator circuit is not constant but will decrease with increasing m.

The width ($t_g$) of the gating pulse is set by means of the aperture-duration controls and circuit, and the delay between receiving a trigger and sampling the following sweep is adjusted by means of the aperture-delay circuit. If the aperture delay is set manually to a constant value, then the boxcar is in a

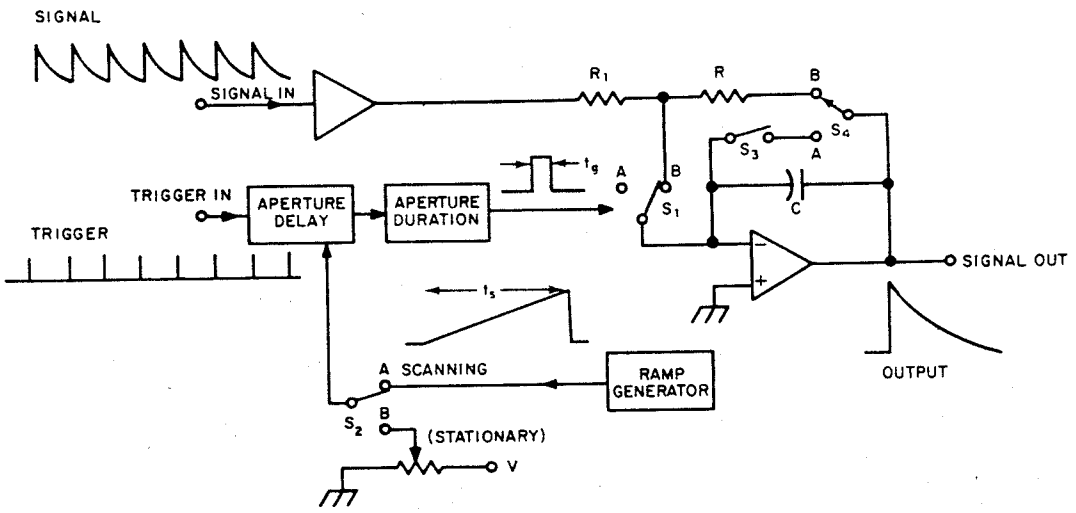



Figure 11.26 The boxcar averager (simplified)

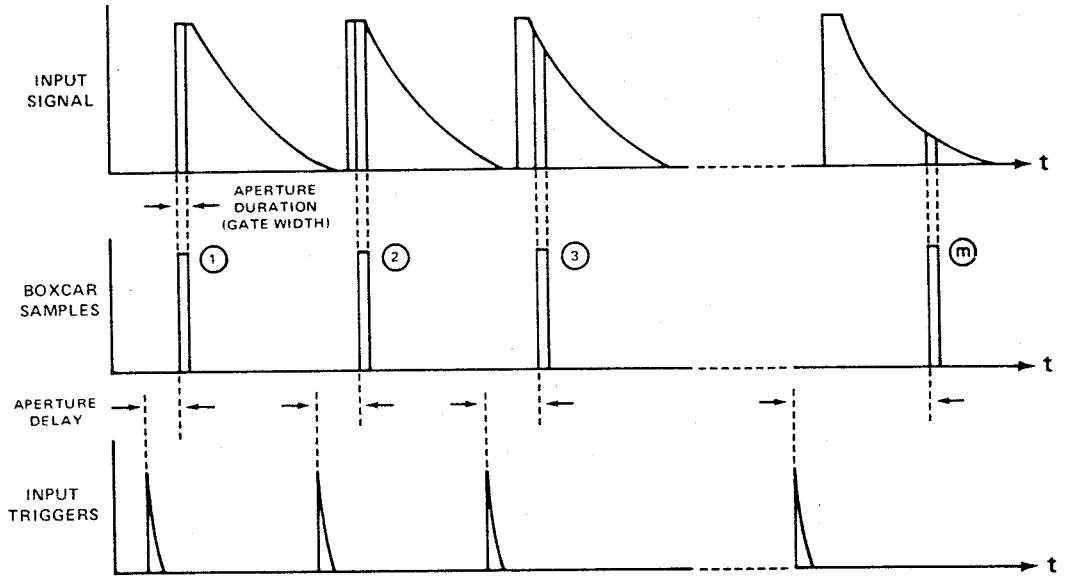



Figure 11.27 Boxar operation in scanning mode (reproduced by permission of EG&G Princeton Applied Research Corporation)

*stationary mode* and will sample the same portion of each successive signal—thus providing an output corresponding to the amplitude of the signal at that point. Alternatively, in its *scanning mode*, the aperture delay can be slowly and continuously changed by a voltage ramp from the scan ramp generator, so that the sampling aperture is slowly moved across the entire signal (see Figure 11.27). In this mode, the boxcar output is a replica of the signal waveform and the boxcar can be regarded as a time-translation device that can slow down and recover fast waveforms.

In the scanning mode, the aperture duration ($t_g$) is not necessarily equal to the time resolution but rather sets the maximum resolution that can be achieved (assuming no input bandwidth limitation) if the scan is sufficiently slow. For an amplitude resolution of within 1 % of the full-scale value, a scan time $T_s$, a signal (sweep) duration of $T$, and a total effective boxcar time constant of $\tau_B$, the time resolution $t_R$, is given by

$$t_R = 5\tau_B T/T_S \tag{11.23}$$

where

$$\tau_B \simeq \sqrt{(\tau_{eff}^2 + \tau_f^2)} \tag{11.24}$$

and $\tau_f$ is the time-constant of any additional filtering used in the instrument.

Boxcar averagers can resolve very fast waveforms. A 100 ps dual-channel boxcar averager using alternate signal sampling and baseline sampling in each channel, is shown in Figure 11.28. Without its averaging capability, a boxcar is

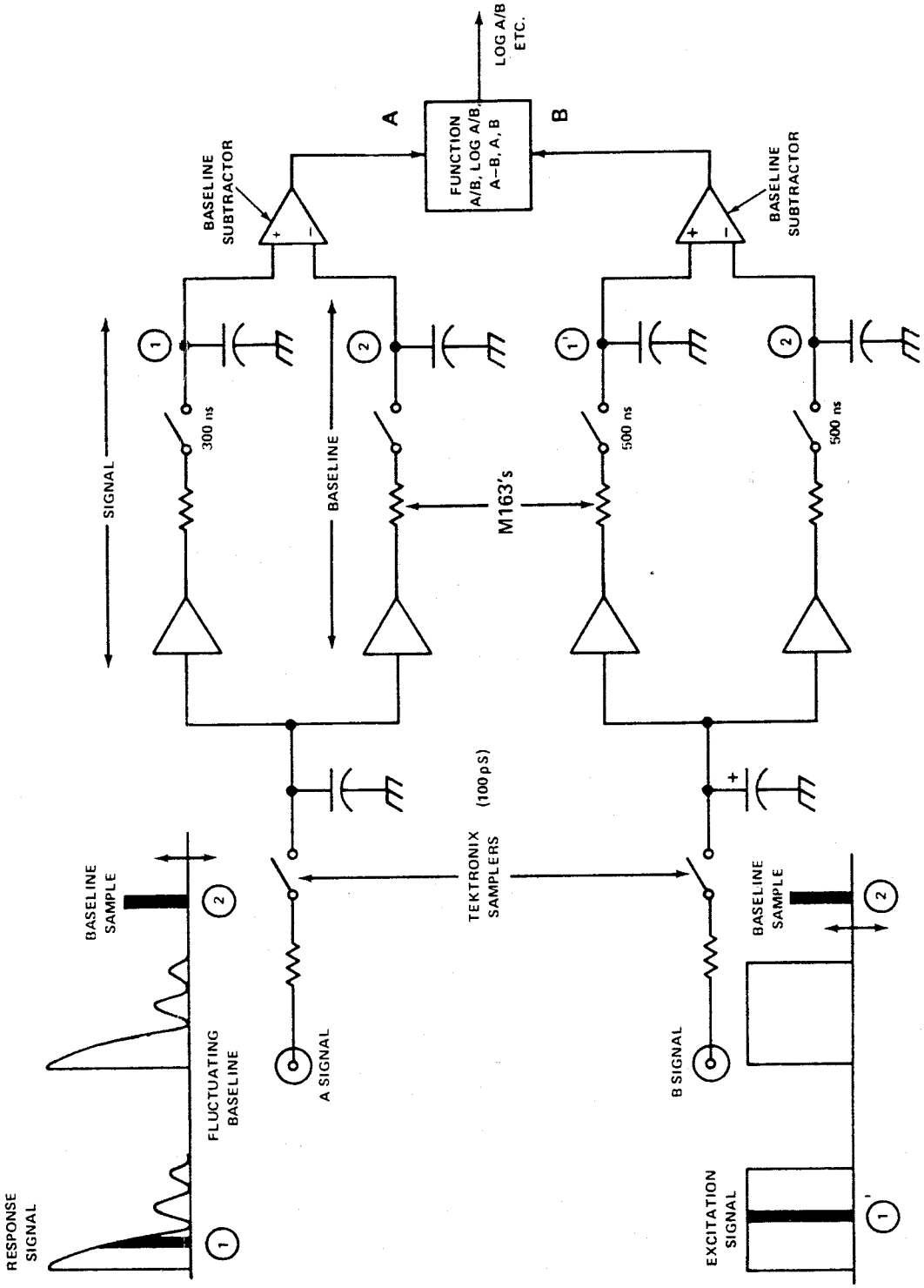



Figure 11.28 Dual channel, 100 ps boxcar system (reproduced by permission of EG&G Applied Research Corporation)

similar to a sampling oscilloscope and as shown in this figure, sample-and-hold plug-ins designed for use in such oscilloscopes can be used as fast front-ends for a boxcar.

### 11.8.2 The Multipoint Signal Averager

The boxcar is a *single-point* averager: it samples each signal occurrence (sweep) only once. A *multipoint averager* (Hewlett Packard, 1968) acts much like a large number of boxcars connected in parallel, since it samples many points (typically $2^{10} = 1024$) during each signal sweep. In such multipoint instruments, the analog storage capacitor of the boxcar is replaced by digital memory; each sample is digitized and the new data are added to the data from previous sweeps already in the memory location corresponding to that sampling point.

Figure 11.29 illustrates some typical waveforms and timing details for a multipoint averager; for simplicity $I = 10$ (i.e. only ten samples/sweep are shown). The total signal duration ($T$) is given by the product of the number of samples/sweep ($I$) and the dwell-time (gate width or sampling duration, $t_g$) of each sample. Notice that $T$ is less than the total sweep duration ($\tau$) by the dead time ($t_d$), and that there is usually a fixed delay time between receipt of a trigger pulse and the beginning of the first sample. Although in most applications a multipoint averager is triggered at a constant rate ($f = 1/\tau$), it is not necessary that the trigger be periodic. Assume that the averager is set to continue averaging until m input sweeps have been sampled, at which point it will automatically stop.

Suppose we wish to recover the waveform of a noisy signal, $f(t)$, where

$$f(t) = s(t) + n(t)$$

For the $i$th sample of the $k$th sweep,

$$f(t) = f(t_k + it_g) = s(t_k + it_g) + n(t_k + it_g) \tag{11.25}$$

For any particular sample point ($i$), the input signal can be assumed to remain unchanged with each new value of $k$ (i.e. with each new sweep) and the averaged signal will therefore be simply

$$S(i)_{out} = \sum_{k=1}^{m} s(t_k + it_g) = ms(it_g) \tag{11.26}$$

For random noise, samples ($X_i$) will add vectorially, so that the r.m.s. value ($\sigma$) of the averaged noise will be given by

$$(\sum (X_i)^2)^{1/2} = \sigma\sqrt{m} \tag{11.27}$$

The averager output can be described by

$$g(t_k + it_g) = ms(it_g) + \sigma\sqrt{m} \tag{11.28}$$

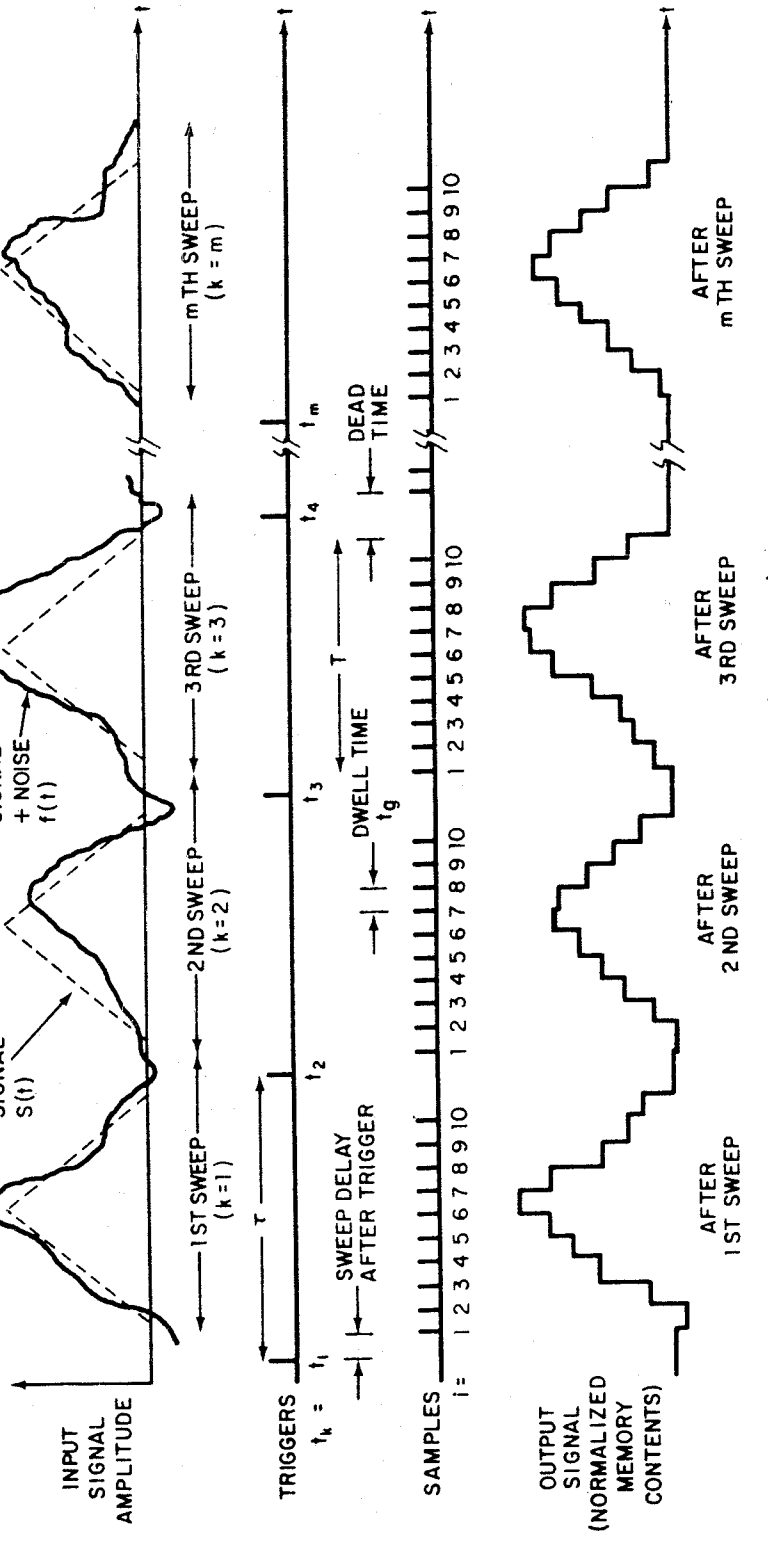



Figure 11.29 Multipoint averaging

so that the output SNR is

$$\mathrm{SNR}_{\mathrm{out}} = \frac{S_{\mathrm{out}}}{N_{\mathrm{out}}} = \frac{ms}{\sigma\sqrt{m}}\, it_{\mathrm{g}} = \frac{s}{\sigma}\, it_{\mathrm{g}}\sqrt{m} \tag{11.29}$$

The input SNR is simply

$$\mathrm{SNR}_{\mathrm{in}} = s(it_{\mathrm{g}})/\sigma \tag{11.30}$$

so that

$$\mathrm{SNIR} = \frac{\mathrm{SNR}_{\mathrm{out}}}{\mathrm{SNR}_{\mathrm{in}}} = \sqrt{m} \tag{11.31}$$

In order to consider a multipoint averager from a frequency domain or filtering point of view, we need to know its transfer function: $H(\mathrm{j}\omega)$. We can determine $H(\mathrm{j}\omega)$ if we know the impulse response $h(t)$, since $H(\mathrm{j}\omega)$ and $h(t)$ are a Fourier-transform pair.

We can determine $h(t)$ heuristically by the following reasoning. In a multipoint averager, trigger pulses are used to synchronize the signal sweeps and allow the signal samples to be coherently added (CO-ADDed). Mathematically, this action can be thought of as convolving the input signal, $f(t)$, with a train of $m$ unit impulses (triggers) spaced $\tau$ seconds apart. The averager's effective impulse response is, therefore, given by

$$h(t) = \sum_{k=1}^{m} \delta(t - k\tau) \tag{11.32}$$

By Fourier transforming this expression for $h(t)$, we find (Childers and Durling, 1975) that the averager's transfer function is

$$|H(\mathrm{j}\omega)| = \left|\frac{\sin(m\omega\tau/2)}{\sin(\omega\tau/2)}\right| \tag{11.33}$$

Notice (from L'Hôpital's rule) that $H(\mathrm{j}\omega) = m$ whenever $\omega t$ is an integral multiple of $2\pi$. Figure 11.30 shows the *comb filter* response of equation (11.33) for several values of $m$. Each band-pass response is centred at a harmonic $(n/\tau)$ of the sweep/trigger rate. (If the trigger rate is aperiodic, then this comb-filter concept becomes meaningless.) Since the peak transmission of each bandpass response is $m$, the $-3$ dB points must occur at $m/\sqrt{2}$, so that

$$|H(\mathrm{j}\omega)| = \left|\frac{\sin(m\omega\tau/2)}{\sin(\omega\tau/2)}\right| = \frac{m}{\sqrt{2}} \tag{11.34}$$

from which the $-3$ dB bandwidth $B$ for large values of $m$, is found to be

$$B = 0.886/(m\tau) \tag{11.35}$$

Large values of $m$ are practicable, particularly at high sweep rates. With a trigger rate $(1/\tau)$ of 100 Hz and $m = 10^6$ for example, the total measurement time will be $m\tau = 10^6 \times 10^{-2} = 10^4$ s $\simeq 2.8$ h, and $B = 8.86 \times 10^{-5}$ Hz.

Thus far in this discussion of multipoint averagers, a *linear summation* mode of averaging has been assumed. That is, for the $i$th memory location, the average after $m$ sweeps is given by

$$A_m = \sum_{k=1}^{m} f(t_k + it_g) = \sum_{k=1}^{m} I_k \tag{11.36}$$

where $I_k = f(t_k + it_g)$ is the value of the $i$th sample in the $k$th sweep.

This algorithm has the advantage of being simple to implement digitally. The output averaged signal, however, continually increases with each new sweep; manual scale changing is required to keep the displayed output at a useful size, yet within the bounds of the CRT screen. A seemingly more convenient algorithm would be to normalize the data in memory after each sweep, that is, implement

$$A_k = \frac{1}{k} \sum_{k=1}^{m} I_k = A_{k-1} + \frac{I_k - A_{k-1}}{k} \tag{11.37}$$

During each sweep, the data ($A_{k-1}$) in each memory location are compared with the new sample value $I_k$ and the computed value of $(I_k - A_{k-1})/k$ is added to memory to form the new average value $A_k$. Because of practical difficulties in implementing a division by $k$ during or after each sweep, the algorithm shown in equation (11.37) is often approximated by

$$A_k = A_{k-1} + \frac{I_k - A_{k-1}}{2^J} \tag{11.38}$$

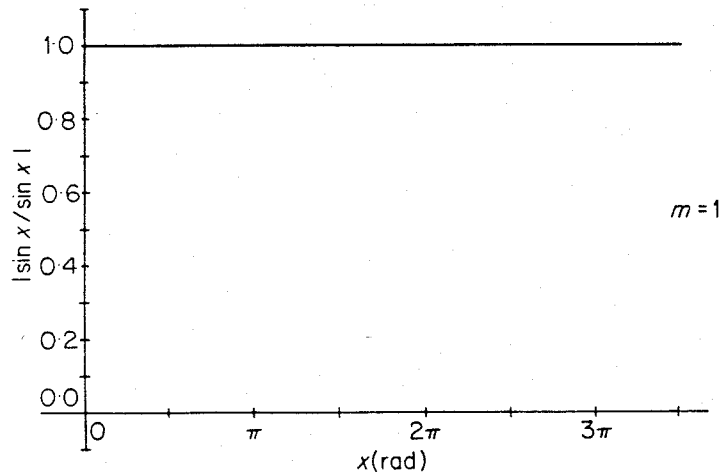
where $J$ is a positive integer selected automatically such that $2^J$ is the closest approximation to $k$. Notice for $k = 6$ for example, that the closest $2^J$ values are $2^2 = 4$ or $2^3 = 8$. Though this *normalized averaging* mode is very slightly slower than the summation mode in enhancing the signal, we can assume that SNIR = $\sqrt{m}$ for all practical purposes. Note that the discrepancy between $k$ and $2^J$ increases as larger values of $J$ are used to deal with very noisy signals. In compensation, this averaging mode provides a stable, constant-amplitude display from which the noise appears to shrink with time.

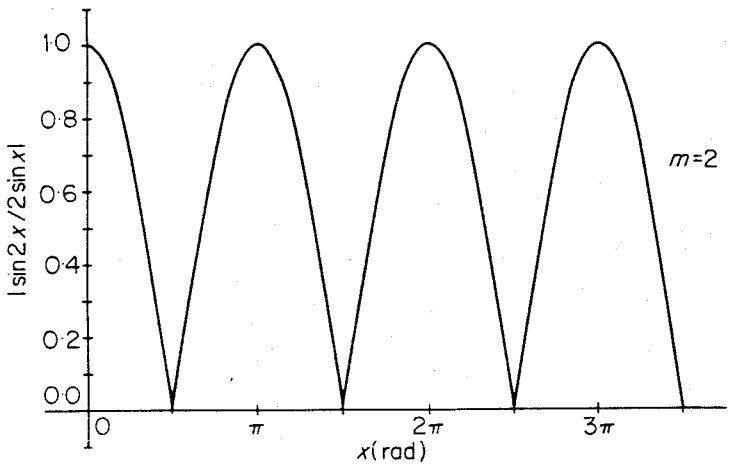
If we wish to recover and monitor slowly varying noisy signals, the algorithm of equation (11.38) can also be used for *exponential* averaging if $J$ is made a manually selectable constant. When $J = 0$, then $2^J = 1$ and $A_k = I_k$; with this setting, the input signal may be monitored in real time, since it is digitized and stored without averaging. In general, selecting a value of $J$ will establish an effective time constant, $\tau_J$, where

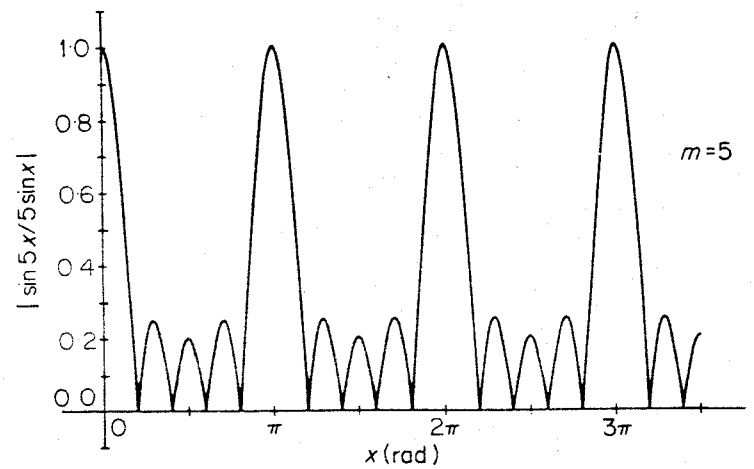
$$\tau_J = \frac{t_g}{-\ln(1 - 2^{-J})} \tag{11.39}$$

or

$$2^{-J} = 1 - \exp(-t_g/\tau_J) \tag{11.40}$$


1·0
0·8
0·6
0·4
0·2
0·0
|sin x/sin x|
m=1
0
π
2π
3π
x(rad)


1·0
0·8
0·6
0·4
0·2
0·0
|sin 2x/2 sin x|
m=2
0
π
2π
3π
x(rad)


1·0
0·8
0·6
0·4
0·2
0·0
|sin 5x/5 sin x|
m=5
0
π
2π
3π
x(rad)

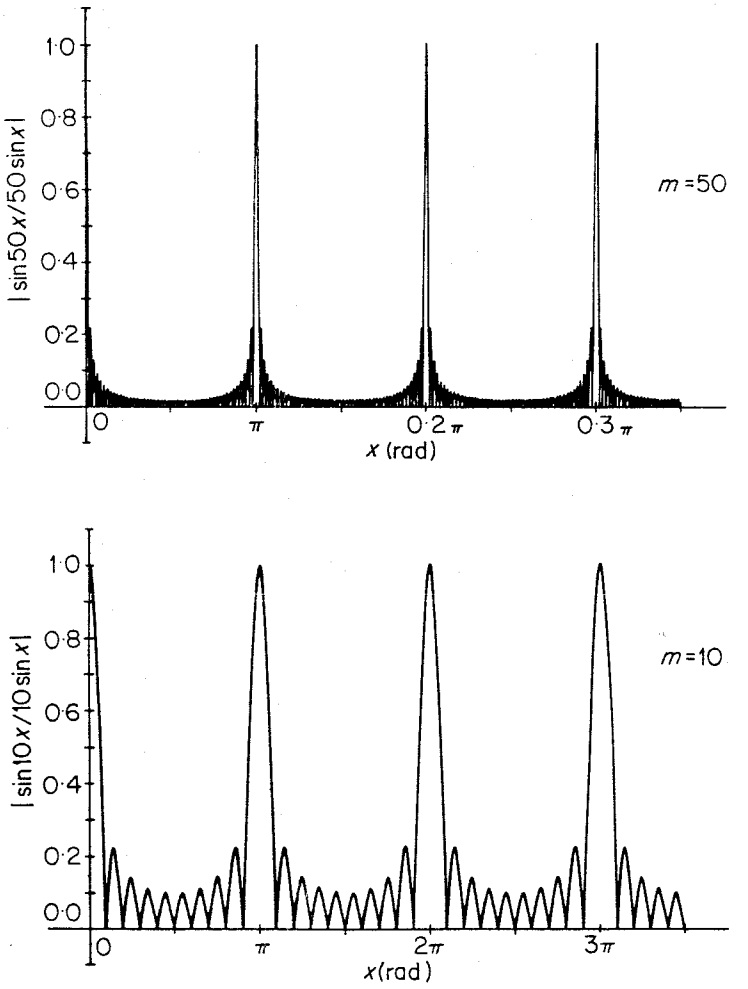



Figure 11.30 The comb filter action of a multipoint averager

The larger the value of $J$ selected, the greater the signal enhancement, and the more slowly the averager responds to changes in the input signal. For a large number of sweeps (EG & G, 3) the SNIR is given by

$$\text{SNIR} \simeq \sqrt{2^{J+1}} \tag{11.41}$$

Figure 11.31 shows the simplified block diagram of a multipoint averager. It is common to include a low-pass filter in the analog input channel with a $-3$ dB cut-off frequency ($f_c$) controlled by the dwell-time setting. A typical example might be

$$f_c \simeq 1/(2t_g)$$

that is, one-half of the sampling frequency. Such filters are used to improve the input SNR rather than as anti-aliasing filters.

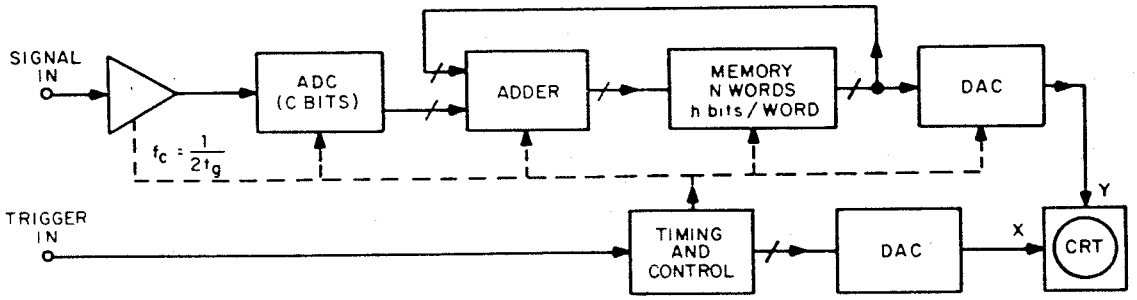



Figure 11.31 The multipoint signal averager (simplified)

The maximum number of sweeps ($m_{max}$) that can be digitized by an averager without data overflow, if the input signal is full-scale and noise-free, is given by $2^{h-c}$ where $h$ is the memory size (bits) and $c$ (bits) is the resolution of the A/D converter (ADC). For a 9-bit (8-bit + sign) ADC and a memory of $N$ words, each of 28 bits, then

$$m_{max} = 2^{h-c} = 2^{28}/2^9 = 2^{19} = 524{,}288$$

In some instruments with *artifact-rejection* capability, each new signal sweep is digitized and placed in a buffer memory. Before adding the buffer contents to main memory, each buffer-memory location is checked for overflow; the buffer contents are discarded should an overflow (artifact) exist.

Suppose that the input SNR to an averager is 1 : 10; that is, the r.m.s. noise ($\sigma$) is ten times larger than the peak signal ($S$). The a.c. gain before the ADC, must be set such that the noise peaks do not exceed full scale. For Gaussian noise, it is 99.9% probable that the peak noise ($N_p$) amplitude is less than five times greater than the rms noise amplitude; that is, $N_p/\sigma \leqslant 5$, so that $N_p/S \leqslant 50$.

Assume that the input gain is set such that $N_p$ is just equal (say) to the full-scale input level of a 9-bit ADC. Assume also that the resolution of the ADC is $2^9$ (= 512), and the memory size $h = 2^{28}$, as before. Of these 9 bits, 6 bits (= $2^6$ = 64) will be required as dynamic reserve (i.e. to handle the input noise), and only 3 bits (= $2^3$ = 8) will be left to quantize the signal ($S$). In this example, then, the maximum number of sweeps before overflow would be

$$m_{max} = 2^{28}/2^3 = 2^{25} \simeq 3.4 \times 10^7$$

Under the conditions of this example, the output (vertical) resolution will not be limited to 3 bits. Random noise accompanying the signal will *dither* the ADC; that is, the noise will modulate the quantization levels of the ADC so as to provide a resolution that increases as $m$ increases. Note, however, that without noise and with the same full-scale setting, the averager output would indeed have a 3-bit amplitude resolution. White noise can be added deliberately to signals that are clean, in order to improve resolution beyond that of the ADC (Horlick, 1975).

It is useful to compare boxcar and multipoint averagers. For dwell times of about 1 $\mu$s or longer, the multipoint averager typically needs less than one-thousandth of the measurement time needed by a boxcar to recover a waveform; on the other hand, the boxcar is the only choice for gate widths (dwell times) of 1 ns or less. For dwell times in the 1 ns–1 $\mu$s range, the choice is between a boxcar or a transient recorder interfaced to a multipoint averager. Such transient recorder–averager combinations are usually less time-efficient (i.e. $\tau \gg It_g$) than a multipoint averager alone, due to slow data transfer.

## 11.9 CORRELATION

For our purposes in this chapter, correlation analysis is a method of detecting any similarity between two time-varying signals (Honeywell). *Autocorrelation* consists of the point-by-point multiplication of a waveform by a delayed or time-shifted version of itself, followed by an integration or summation process. Mathematically, the autocorrelation function, $R_{xx}(\tau)$, of a time-varying function, $f(t)$, is given by

$$R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t)f(t+\tau)\,dt \tag{11.42}$$

where $\tau$ is the *lag value* or time shift between the two versions of $f(t)$.

*Cross-correlation* involves two waveforms and consists of the multiplication of one waveform, $f(t)$, by a time-shifted version of a second waveform, $g(t)$, followed by integration or summation. The cross-correlation function, $R_{xy}(\tau)$, is given by

$$R_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t)g(t+\tau)\,dt \tag{11.43}$$

Notice that cross-correlation requires two input signals, as is also true in the case of signal averaging (where a synchronizing input is required in addition to the signal input). Also as with an averager, a cross-correlator preserves signal phase information. Unlike the averager, however, the cross-correlator output waveform, the *correlogram*, is affected by the waveform of the second input signal—an undesirable and unnecessary complication for signal recovery applications since a multipoint averager may be used. Cross-correlators are normally used in flow or velocity measurements; they are used but rarely for simple signal-recovery purposes. (Ignoring the lock-in amplifier and boxcar integrator, both of which can be regarded as a special type of cross-correlator.)

Phase information is lost in an autocorrelation function, as is also true for its Fourier transform, power spectral density. This lack of phase information means that in some cases, the input signal responsible for a given correlogram must be deduced by intelligent guesswork. For example, as shown in Figure 11.32, the autocorrelation function for band-limited white Gaussian noise is a

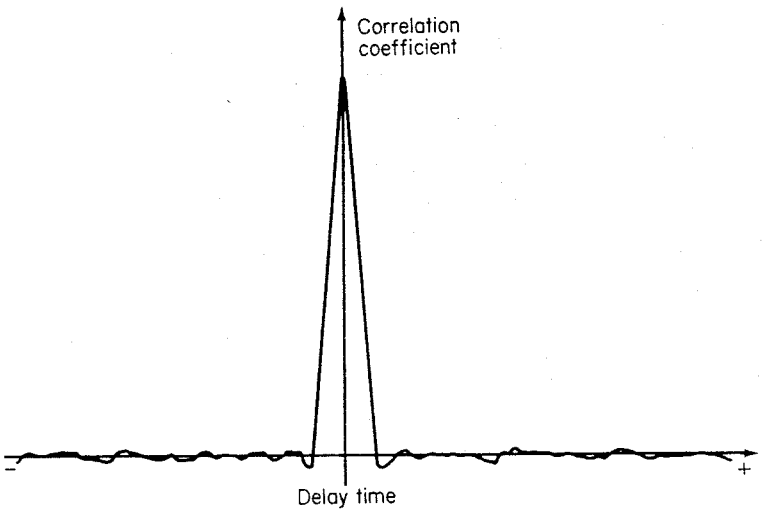



Figure 11.32 Correlation function of band-limited white noise

spike-like peak at $\tau = 0$, with a width that would decrease as the noise bandwidth increases. A similar correlogram could have resulted from an input consisting of a single narrow pulse; the narrower the pulse width, the narrower the correlogram spike. Thus the pulse and the band-limited white noise have similar power-density spectra. (The difference between them is that frequency components of the noise have random phase relationships.)

A simplified block diagram of an autocorrelator is shown in Figure 11.33. The ADC will digitize the input signal once every lag interval or dwell time, $t_g$. Each such A/D conversion will require a conversion time, $t_c$, where $t_c \ll t_g$. The output digital *word* from the ADC, corresponding to the latest sample, provides one input to the digital multiplier and also is shifted, as word 0, into an $N$-word shift register. During this shift operation, the last word in the register, word $(N - 1)$, is shifted out and discarded and the former word $(N - 2)$ becomes the new word $(N - 1)$. The control and timing circuits

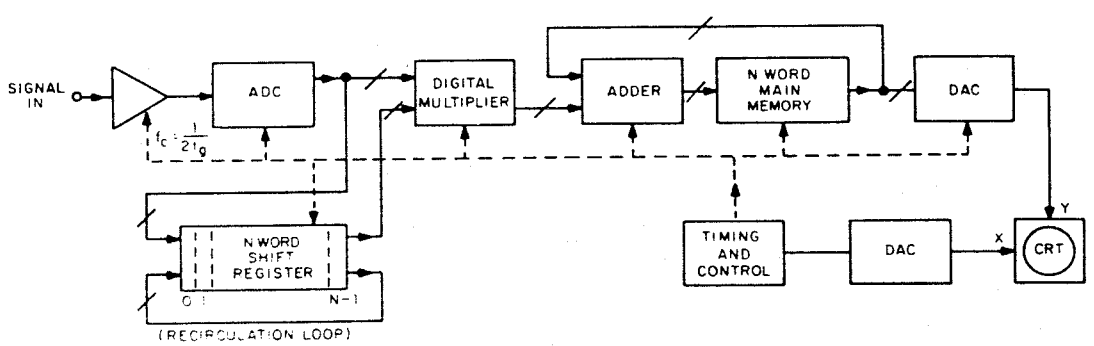



Figure 11.33 The autocorrelator (simplified)

then cause the shift register to step $N$ times, recirculating its contents one full cycle, and requiring a time interval $t_r$. During $t_r$, each shift-register word (i.e. $(n-1), (n-2), (n-3), \ldots, 3, 2, 1, 0$) is applied sequentially to the other input of the multiplier, which multiplies each of these words by the word at its other input, word 0. Each multiplier output is added to the contents of the corresponding bin of the $N$-bin main memory; for instance, word 0 $\times$ word $(n-16) \rightarrow$ bin$(n-16)$, word 0 $\times$ word 0 $\rightarrow$ bin 0. After many such cycles, the contents of bin $i$ (for example) of the main memory, will be the sum of products formed by multiplying each new signal sample by an $it_g$ delayed version of itself. Bin 0, for example, corresponds to the signal multiplied by itself with zero delay.

The minimum time between successive samples is $(t_c + t_r)$. When $t_g \geqslant (t_c + t_r)$, the correlator is said to be working in a *real-time* mode. When $t_g \leqslant (t_c + t_r)$, samples can no longer be taken every $t_g$ seconds and the correlator is said to be in a *pseudo-real-time* mode. For $t_g \ll (t_c + t_r)$, the correlator is in a *batch* mode.

The process of autocorrelation involves the concept of sliding a waveform past a replica of itself. For random noise, the two waveforms will match at only one point as they slide by each other, that is at $\tau = 0$, when they are perfectly aligned. In contrast, a square-wave sliding by another square-wave will find a perfect match once in every period and will give rise to a triangular correlogram. More generally, signals that are periodic in time will produce a correlation function that is periodic in $\tau$. Suppose, for example, that the input to an autocorrelator is $f(t) = A\cos(\omega t)$. Then

$$R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} A\cos(\omega t) A\cos(\omega t + \tau)\,dt = \tfrac{1}{2}A^2\cos(\omega\tau) \quad (11.44)$$

Figure 11.34 shows the correlation function of a sine wave accompanied by band-limited white noise. Note that the peak value at $\tau = 0$ in this correlogram is the mean-squared value of the signal plus noise (i.e., $S + N$). The mean-squared value of the sinusoidal signal component is given by the peak value of

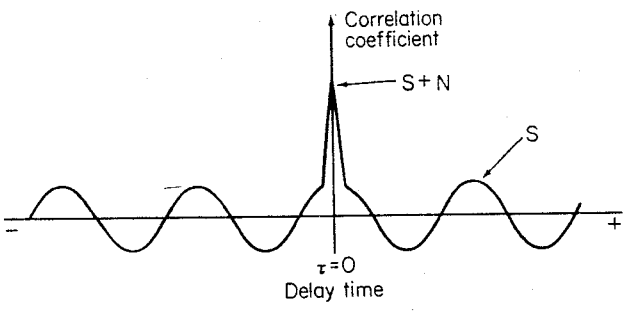



Figure 11.34 Correlogram of a noise sinewave

the sine wave at $\tau$ values where the white noise spike has damped to zero. The SNR of the correlator output can, therefore, be determined since

$$\frac{S}{(S+N)-S} = \frac{S}{N} = \text{SNR}$$

Most importantly, notice that the signal and noise components have been separated by virtue of their different positions on the $\tau$-axis. It is this separating ability that makes correlation a powerful signal-recovery technique.

## 11.10 PHOTON (PULSE) COUNTING TECHNIQUES

### 11.10.1 Introduction

PMT's (photomultiplier tubes) are used to measure the intensity or flux of a beam of visible photons (see Figure 11.35). With its photo-cathode removed, the PMT becomes an EMT (electron multiplier tube) and is widely used to detect ions and electrons. One of the most important advantages of such detectors is that their high gain and low noise allow them to give one output pulse for each detected input particle. Since visible-light measurements are perhaps most commonplace, we will consider as an example a PMT and pulse-counting system of the type shown in Figure 11.36.

The probability of each incident photon causing an output pulse from the PMT is essentially equal to the *quantum efficiency*, $\zeta$: typically $\zeta$ is between 5–25%. In addition to the photon-derived pulses (i.e. the signal pulses), there will be spurious noise pulses at the PMT output due to thermionic emission.

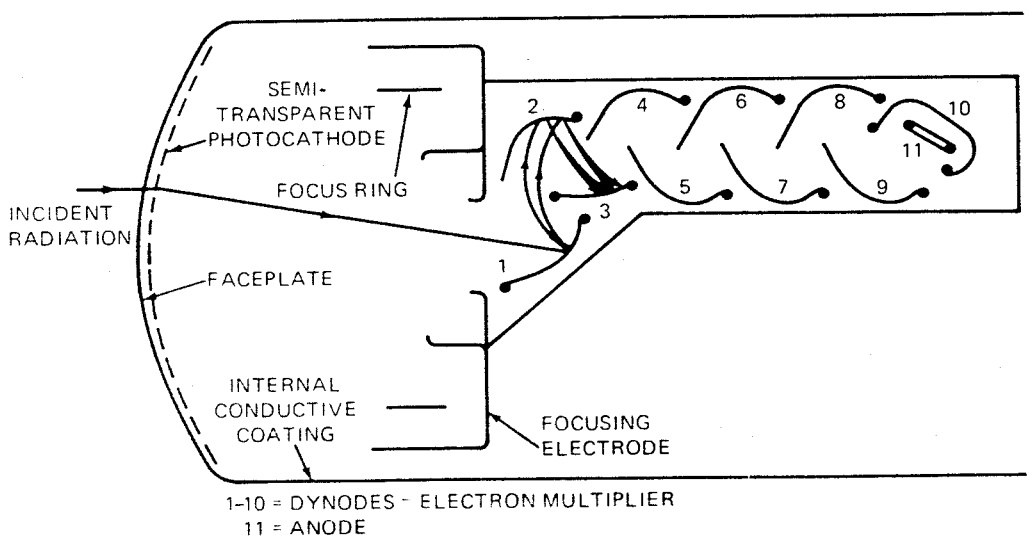



Figure 11.35 End-window photomultiplier tube

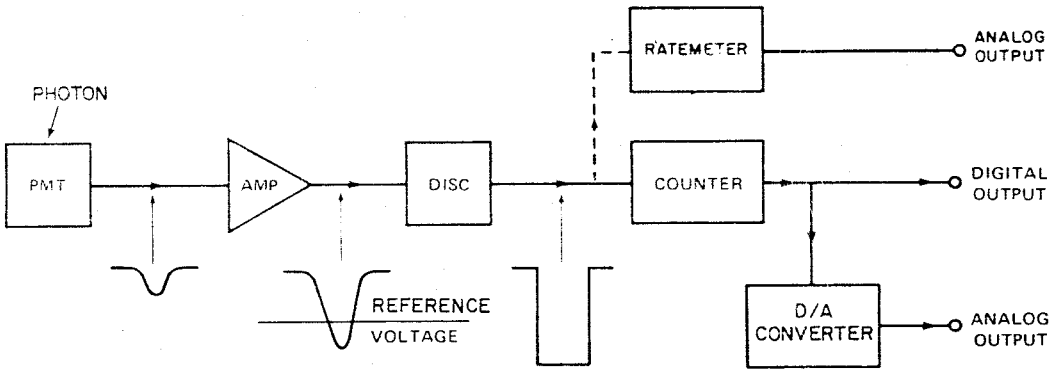



Figure 11.36 Typical photon counting system (reproduced by permission of EG&G Princeton Applied research Corporation)

Noise pulses caused by thermionic emission from the dynodes will experience less gain and will be smaller in amplitude than pulses due to cathode emission. The PMT output pulses are amplified and presented to a pulse-height discriminator circuit, where the peak amplitude of each pulse is compared to an adjustable threshold or reference voltage. Ideally, the discriminator will reject all dynode-derived noise pulses and accept all cathode-derived pulses; in practice, the PMT gain is statistical in nature and cathode and dynode-derived pulses have overlapping amplitude distributions. The discriminator will therefore accept *most* cathode-derived pulses and reject *most* dynode noise pulses. Each accepted input pulse will cause a standardized output pulse. Such pulse-height discrimination also reduces the effect of PMT gain variations with time and temperature.

Ratemeters are used to give a continuous analog output voltage which is proportional to the discriminator output pulse (count) rate. Alternatively, digital counter circuits can be used to accumulate output counts for a preselected measurement time. Such counters allow very long integration times and when a digital output is required, they can avoid the loss of resolution inherent in D/A conversion.

### 11.10.2 Poisson Statistics, Shot Noise, and Dark Counts

Suppose we use our PMT to detect photons emitted from a thermal light source such as a tungsten filament lamp. The time interval between successive photons impinging upon the PMT photocathode is random and governed by a *Poisson* distribution (see Figure 11.37a). The probability, $P$, of detecting $n$ photons in a time $t$ following the last photon is given by

$$P(n, t) = \frac{(\zeta R t)^n e^{-\zeta R t}}{n!} = \frac{N^n e^{-N}}{n!} \tag{11.45}$$

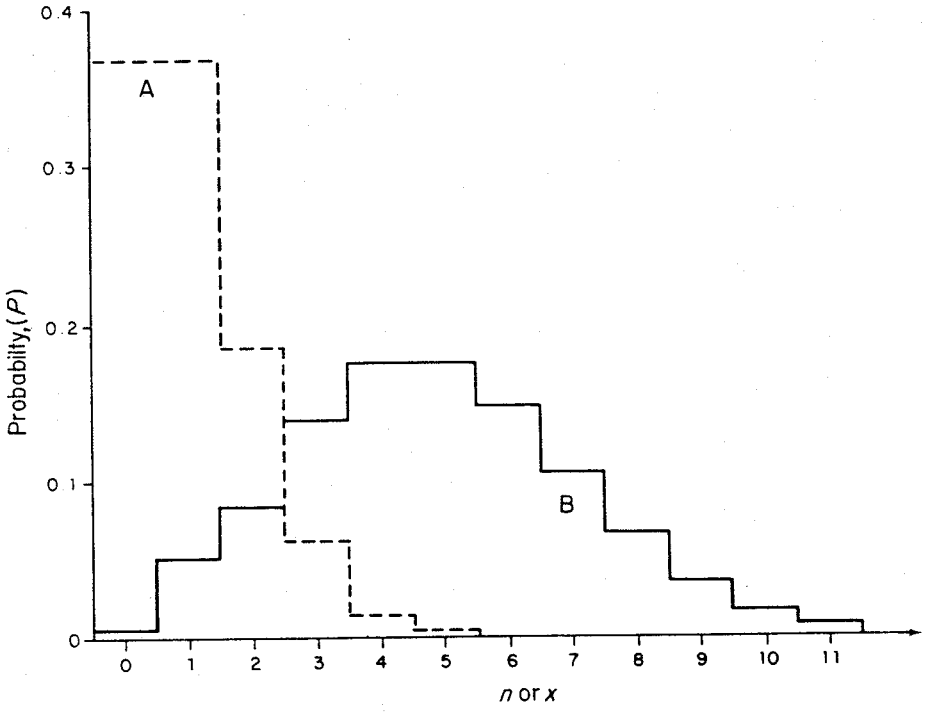



Figure 11.37 The Poisson distribution (reproduced by permission of EG&G Princeton Applied Research Corporation). Curve A, Probability of detecting $n$ photons in time $t = (Rt)^n \exp(-Rt/n)$, $R = 10^8$ photons/s, $t = 10$ ns and so $Rt = 1$; $\sigma = \sqrt{(Rt)}$. Curve B, Probability of gain magnitude $P(x) = M^x e^{-M}/x!$, where $x$ = dynode gain, $M$ = mean dynode gain = 5 and $\sigma = \sqrt{M}$.

where $R$ is the mean photon rate (photons/second) and $N = \zeta Rt$ is the signal (i.e. the mean number of photonelectrons emitted by the PMT photocathode during the time interval $t$). The noise, or uncertainty, in $N$ is given by the standard deviation $\sigma$, where

$$\sigma = \sqrt{(\zeta Rt)} = \sqrt{N}$$

so that

$$\text{SNR} = \frac{N}{\sqrt{N}} = \sqrt{N} = \sqrt{(\zeta Rt)} \tag{11.46}$$

Notice that, as in all of the techniques examined in this chapter (with white noise), the SNR is again proportional to the square root of the measurement time ($t$). If we assume that there is no thermionic (dark) emission of electrons from the photocathode, then the photocathode (signal) current (in amperes) is given by

$$I_{pe} = \zeta Re \tag{11.47}$$

where $e$ is the charge of an electron ($\simeq 1.6 \times 10^{-19}$ C). The signal-to-noise ratio ($SNR_k$) of the photocathode current ($I_{pe}$) is given by

$$SNR_k = \sqrt{(\zeta Rt)} = \sqrt{\left(\frac{\zeta Rte}{e}\right)} = \sqrt{\left(\frac{I_{pe}t}{e}\right)} \tag{11.48}$$

Now the measurement time $t$ has associated with it a frequency range $\Delta f$, where

$$t = 1/2\Delta f,$$

so that

$$SNR_k = \sqrt{\left(\frac{I_{pe}}{2e\Delta f}\right)} = \sqrt{\left(\frac{I_{pe}^2}{2eI_{pe}\Delta f}\right)} = \frac{I_{pe}}{\sqrt{(2eI_{pe}\Delta f)}} \tag{11.49}$$

If we multiply both the numerator and denominator of equation (11.49) by the mean PMT gain $A$, then

$$SNR_k = \frac{AI_{pe}}{A\sqrt{(2eI_{pe}\Delta f)}} = \frac{I_a}{\sqrt{(2AeI_a\Delta f)}} = SNR_a \tag{11.50}$$

where $I_a$ is the d.c. anode current and $SNR_a$ is the signal-to-noise ratio of the anode current if thermionic emission and other dynode noise contributions are ignored. Note that the general expression for the shot noise of a d.c. current $I$ is given by

$$\text{r.m.s. shot noise current} = \sqrt{(2AeI\Delta f)} \tag{11.51}$$

where $A$ is the gain following the shot noise process. When $A = 1$, the expression simplifies to $\sqrt{(2eI\Delta f)}$. Notice also that shot noise was present in the light beam itself and that the PMT quantum efficiency ($\zeta$) degrades the SNR by a factor of $\sqrt{\zeta}$.

In practice, with no input photons, the photocathode will emit electrons due to temperature effects. The dynodes will also emit thermionic electrons. The rate of such thermionic emission is reduced by cooling the PMT. Thermionic emission from the photocathode, that is, *dark counts*, can be further reduced by minimizing the cathode area and by selecting a photocathode material with no more red (long-wavelength) spectral response than is necessary.

If the photocathode emits electrons randomly at a dark count rate $R_d$, then the noise components of the cathode current will increase to $\sqrt{(\zeta Rt + R_d t)}$ and the signal-to-noise ratio of the cathode current will degrade to

$$SNR = \frac{\zeta Rt}{\sqrt{(\zeta Rt + R_d t)}} = \frac{\zeta R\sqrt{t}}{\sqrt{(\zeta R + R_d)}} \tag{11.50}$$

This will also be the PMT output SNR if dynode noise is assumed to be removed completely by pulse-height discrimination. For PMT's equipped

with a high-gain first dynode with Poissonian statistics (see Figure 11.37b), this is a not-unreasonable approximation. Note that when a PMT is used in a non-counting or d.c. mode, as was discussed previously in Section 11.6, all of the output electrons resulting from spurious cathode emission (i.e. dark counts) and dynode emission are integrated by the anode or preamplifier time constant into a d.c. *dark current*, and the opportunity to remove dynode noise by pulse-height discrimination is lost.

### 11.10.3 Pulse-height Discrimination

Each electron emitted by a PMT photocathode will be amplified by the instantaneous value of the PMT gain. For a mean gain of $10^6$, for example, each cathode electron will cause an average output charge $q$ of $10^6 e$ coulomb. This charge $q$ will accumulate at the anode during a time $t$ given by the transit-time spread of the PMT. Typically, $t$ will be about 10 ns, so that the resulting anode current pulse ($i_a = dq/dt$) will have a full width ($t_w$) between half-maximum amplitude points (FWHM) of about 10 ns also. The peak value, $I_{pk}$, of the anode-current pulse may be approximated by assuming the pulse to be rectangular, so that in our example,

$$I_{pk} \simeq \frac{q}{t_w} = \frac{10^6 e}{10 \times 10^{-9}} = \frac{10^6 \times 1.6 \times 10^{-19}}{10 \times 10^{-9}} = 16\ \mu\text{A}$$

In a photon-counting system, the anode-load resistor ($R_a$) of the PMT is kept small, usually 50–100 Ω. Therefore the time constant ($\tau_a$) formed by the anode stray capacitance ($C_a$) will be small compared to $t_w$, and thus will not stretch the anode voltage pulse. Typically, $R_a = 50\ \Omega$ and $C_a = 20$ pF, so that $\tau_a = 1\ \text{ns} \ll t_w$. The anode voltage pulse will then have the same shape as the anode current pulse, and a peak value of

$$E_{pk} = I_{pk} R_a = 16 \times 10^{-6} \times 50 = 0.8\ \text{mV}$$

It should be remembered that such pulse amplitudes depend upon the PMT gain which, in turn, depends upon the dynode gains—which are statistical. In the above example then, $E_{pk} = 0.8$ mV is the average pulse height to be expected. Actual pulse heights will be distributed above and below this value and the better the PMT, the narrower will be this distribution. A preamplifier is normally used to amplify the anode pulses to a suitable level for the pulse-height discriminator.

Notice that the PMT gain, the preamplifier gain, and the discriminator threshold controls may all be used to adjust the effective discrimination level. Figure 11.38 shows a typical count-rate variation with PMT high voltage (i.e. PMT gain) and a fixed discriminator threshold level. This is one of a family of curves that could be plotted for different threshold levels; similar curves

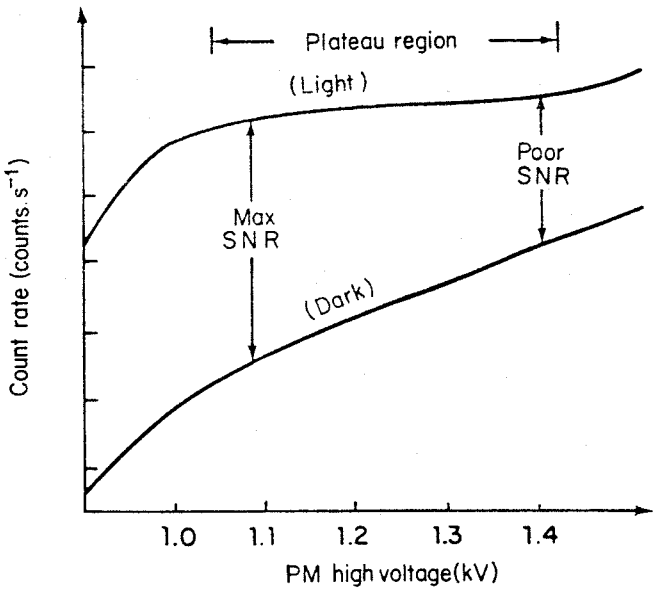



Figure 11.38 The counting plateau (reproduced by permission of EG&G Princeton Applied Research Corporation)

could be obtained by varying the preamplifier gain rather than that of the PMT, or by plotting count rate against discriminator threshold (Darland *et al.*, 1979). The upper curve in Figure 11.38 was plotted by allowing light to fall upon the PMT photocathode and slowly varying the PMT voltage (which is non-linearly related to the PMT gain). Notice that the steep slope at low PMT voltages begins to flatten and form a (not-quite-horizontal) *plateau* as proper focusing takes place in the PMT. The increasing slope at very high PMT bias voltages is due to increasing instability in the PMT.

The upper curve corresponds to $S + N$, since it is based on both signal and noise pulses. The lower curve was plotted with the PMT in darkness, and therefore represents noise ($N$) pulses only. Notice that typically the dark-count curve has no plateau; it has been suggested that the lack of a plateau is due to corona effects associated with microscopic protrusions from the dynode surfaces. Since the count rate is plotted on a logarithmic scale, the vertical distance between the two curves corresponds to

$$\log(S + N) - \log(N) = \log\left(\frac{S + N}{N}\right) = \log\left(1 + \frac{S}{N}\right) \simeq \log\left(\frac{S}{N}\right) \quad \text{for } \frac{S}{N} \gg 1$$

A commonly used method of setting the PMT high voltage for a given preamplifier gain and discriminator threshold level is, therefore, to select a point on the beginning of the counting plateau corresponding to maximum SNR.

### 11.10.4 Ratemeters and Counters

A simplified ratemeter circuit is shown in Figure 11.39. Each output pulse from the discriminator results in a precise current pulse being averaged by the low-pass filter. The ratemeter output voltage is, therefore, proportional to the average value of the discriminator output count rate ($R_{sig}$).

A digital alternative to the ratemeter, a timer–counter circuit, is shown in simplified form in Figure 11.40. Both counters, A and C, are started and stopped together. Counter C is a presettable counter: a number $N$ is preset, usually by means of thumbwheel switches, and the counter will stop when its accumulated count equals $N$. When driven from an internal clock oscillator, this arrangement is called a *timer* circuit. In the *normal* mode, a 1 MHz internal clock is often used so that the timer is used to set the measurement time $t = N/R_{clk}$ μs. The output count will be simply $A = tR_{sig}$.

The ratio mode is usually used for source compensation where the signal count rate is proportional to both the measurand and (say) the intensity of a light source. By monitoring the light source with a separate PMT and amplifier–discriminator to produce a source-dependent count rate $R_{sc}$, $R_{sig}$ can be normalized by $R_{sc}$ to provide an output count that is independent of source fluctuations.

In the *reciprocal* mode, the system will measure the time ($t$, in μs) required for the cumulative signal counts to reach $N$; the smaller the signal count rate, the longer the elapsed time. If the dark count rate is negligible, then the measurement accuracy is $1/\mathrm{SNR} = 1/\sqrt{(R_{sig}t)}$ and for a constant value of $R_{sig}t$ ($= N$), all measurements will have the same SNR and accuracy.

A *synchronous counting system* that acts like a *digital lock-in amplifier* to provide automatic background subtraction is shown in Figure 11.41. When the chopper blade blocks the input light, the output pulses from the amplifier–discriminator are, by definition, background noise; these pulses ($N$) are gated into counter B by the timing circuit—which is itself synchronized by the chopper reference signal. When the chopper blade allows light to reach the PMT, the discriminator output consists of signal-plus-background pulses

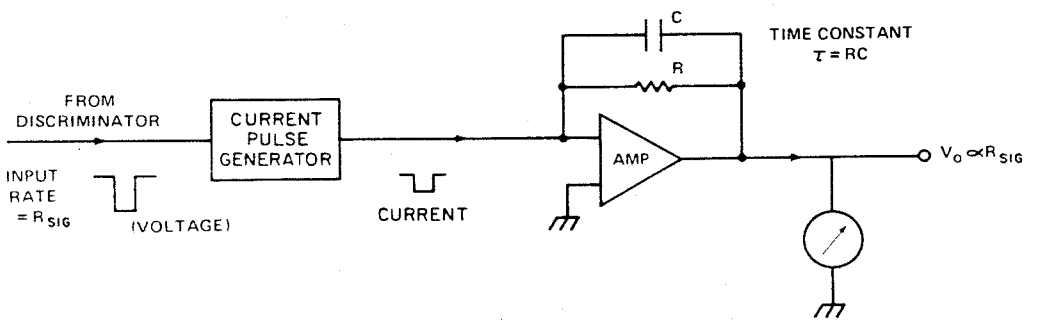



Figure 11.39 The ratemeter (simplified) (reproduced by permission of EG&G Princeton Applied Research Corporation)

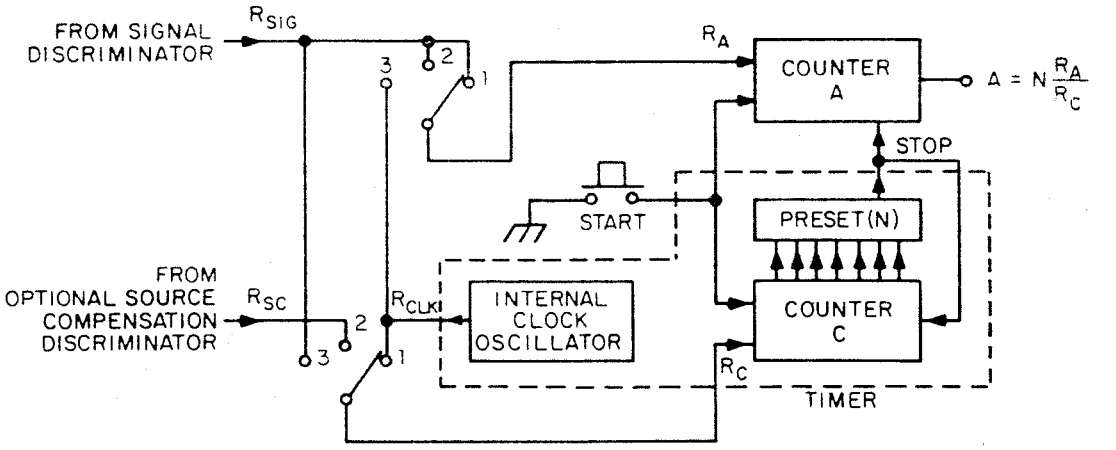



Figure 11.40 Counter–timer modes of operation. (1) Normal mode; $R_A = R_{sig}$, $R_c = R_{clk}$, $A = NR_{sig}/R_{clk} \propto R_{sig}$. (2) Ratio mode; $R_A = R_{sig}$, $R_c = R_{sc}$, $A = NR_{sig}/R_{sc} \propto R_{sig}/R_{sc}$. (3) Reciprocal mode; $R_A = R_{clk}$, $R_c = R_{sig}$, $A = NR_{clk}/R_{sig} \propto 1/R_{sig}$.

$(S + N)$ and these pulses are gated into counter A. After each measurement interval, an arithmetic circuit provides two outputs

$$A - B = (S + N) - N = S = \text{signal} \tag{11.52}$$

and

$$A + B = (S + N) + N = \text{total counts} \tag{11.53}$$

where $A$ and $B$ are the numbers of counts in counters A and B respectively. For Poissonian noise,

$$\text{SNR} = \frac{\text{signal}}{\sqrt{(\text{total counts})}} = \frac{A - B}{\sqrt{(A + B)}} \tag{11.54}$$

Suppose, for example, that $A = 10^6$ counts and $B = 9.99 \times 10^5$ counts then $S = A - B = 10^3$ counts and $\sqrt{(A + B)} = 1.41 \times 10^3$, so that

$$\text{SNR} = \frac{A - B}{\sqrt{(A + B)}} = \frac{10^3}{1.41 \times 10^3} = 0.71$$

and (in)accuracy

$$\frac{1}{\text{SNR}} = \frac{1}{0.71} = 141\%$$

or, expressed in words, the measurement is worthless! The $A + B$ output is important since it allows the measurement accuracy to be estimated in this way.

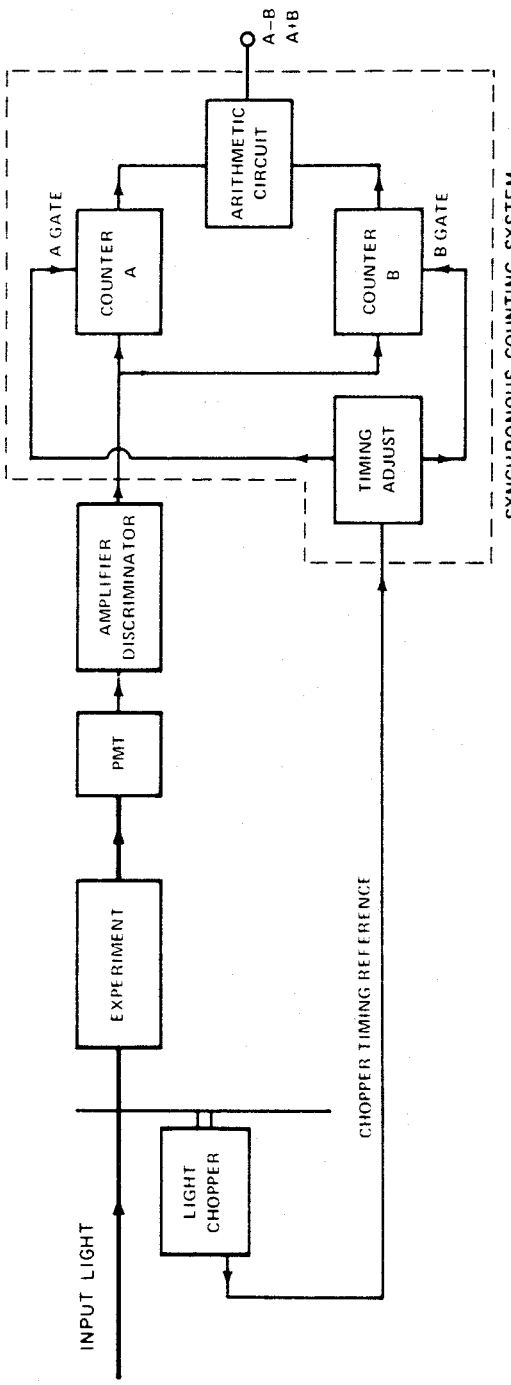



Figure 11.41 Automatic background subtraction (reproduced by permission of EG&G Princeton Applied Research Corporation)

### 11.10.5 Pulse Pile-up

The dynamic range of photon-counting measurements is limited at low light levels by PMT dark count and, at high light levels, by *pulse pile-up* in the PMT or electronics. As the mean rate ($R$) of photons arriving at a PMT photocathode increases, then so does the probability of two or more photons arriving with too short an interval between them to be resolved by the PMT.

The time-resolution of a PMT is effectively equal to its output pulse width $t_w$, and each output pulse from a PMT, therefore, occurs whenever an electron is emitted after a time greater than $t_w$ following the previous electron. The probability of this happening is the same as that for zero photoelectron events in a time $t_w$ (we can neglect dark counts at high light levels). As shown in Figure 11.37a and from equation (11.45) then

$$P(0, t_w) = \exp(-\zeta R t_w) \tag{11.55}$$

and the output count rate, $R_o$, is given by

$$R_o = P(0, t_w)R_i = \zeta R \exp(-\zeta R t_w) = R_i \exp(-R_i t_w) \tag{11.56}$$

The resulting PMT pulse pile-up error is given by

$$\varepsilon_{pmt} = \frac{R_i - R_o}{R_i} = \frac{R_i - R_i \exp(-R_i t_w)}{R_i} = 1 - \exp(-\zeta R t_w) \tag{11.57}$$

The PMT is a *paralysable* detector; that is, when the input count rate exceeds a certain value ($R_i = 1/t_w$), the output count rate will begin to *decrease* for an increasing input count rate and will become zero when the PMT is completely paralysed (saturated) (See Figure 11.42).

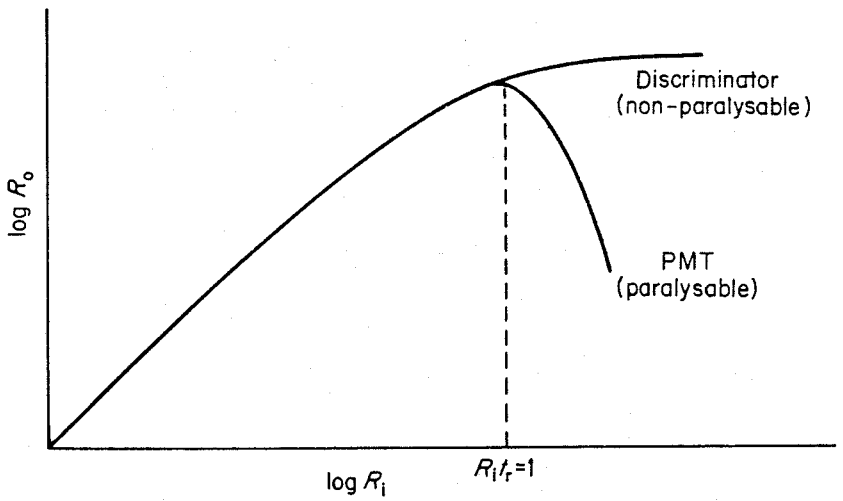



Figure 11.42 Counting Error due to pulse pile-up (reproduced by permission of EG&G Princeton Applied Research Corporation)

Discriminators and counters, on the other hand, are usually non-paralysable. Suppose a discriminator, for example, has a pulse-pair resolution or dead time $t_d$. That is, each time it accepts an input pulse, it cannot accept a new pulse until after a time $t_d$. Then for a measurement time $t$, an input pulse rate of $R_i$ and an output rate $R_o$, the total number of output pulses, $N_o$, is given by

$$N_o = R_o t$$

and

$$\text{the total } \textit{dead} \text{ time} = N_o t_d = R_o t t_d$$

so that

$$\text{total } \textit{live} \text{ time} = t - R_o t t_d$$

The total number of input pulses accepted is therefore given by

$$N_o = R_o t = R_i(t - R_o t t_d)$$

so that

$$R_o = \frac{R_i}{1 + R_i t_d}$$

A modern, fast discriminator and counter have a dead time, $t_d$, of about 10 ns—similar to the time resolution, $t_w$, of a reasonably fast PMT. Notice, however, that any PMT pile-up will act as a *prefilter* to the discriminator; that is, such pile-up will decrease the input count rate to the discriminator. PMT pile-up usually provides the upper limit to the system dynamic range. Photon-counting systems cannot be used in pulsed light measurements where the peak photon rate (during the light pulse) will cause unacceptably high pulse pile-up errors.

## 11.11 FINAL COMMENTS

Many of the signal-recovery considerations discussed in this chapter, such as instrument selection, are summarized in Figure 11.43. Notice that this figure includes two instruments not mentioned in the preceding sections of this chapter: the multichannel analyser (MCA), and the photon (digital) correlator. The choice of signal-recovery instrument is often limited to that which is available and, in some applications, the MCA can be used in place of a multipoint averager. Similarly, a photon correlator, if available, may possibly be substituted for an analog correlator.

In its multichannel scaling (MCS) mode, the MCA (Nicholson, 1974) consists effectively of a scaler (counter) connected to a digital memory much like that of a multipoint averager. During each sweep, the scaler sequentially counts the number of input pulses during each dwell time and adds that number to the cumulative count in the corresponding memory address. By using a

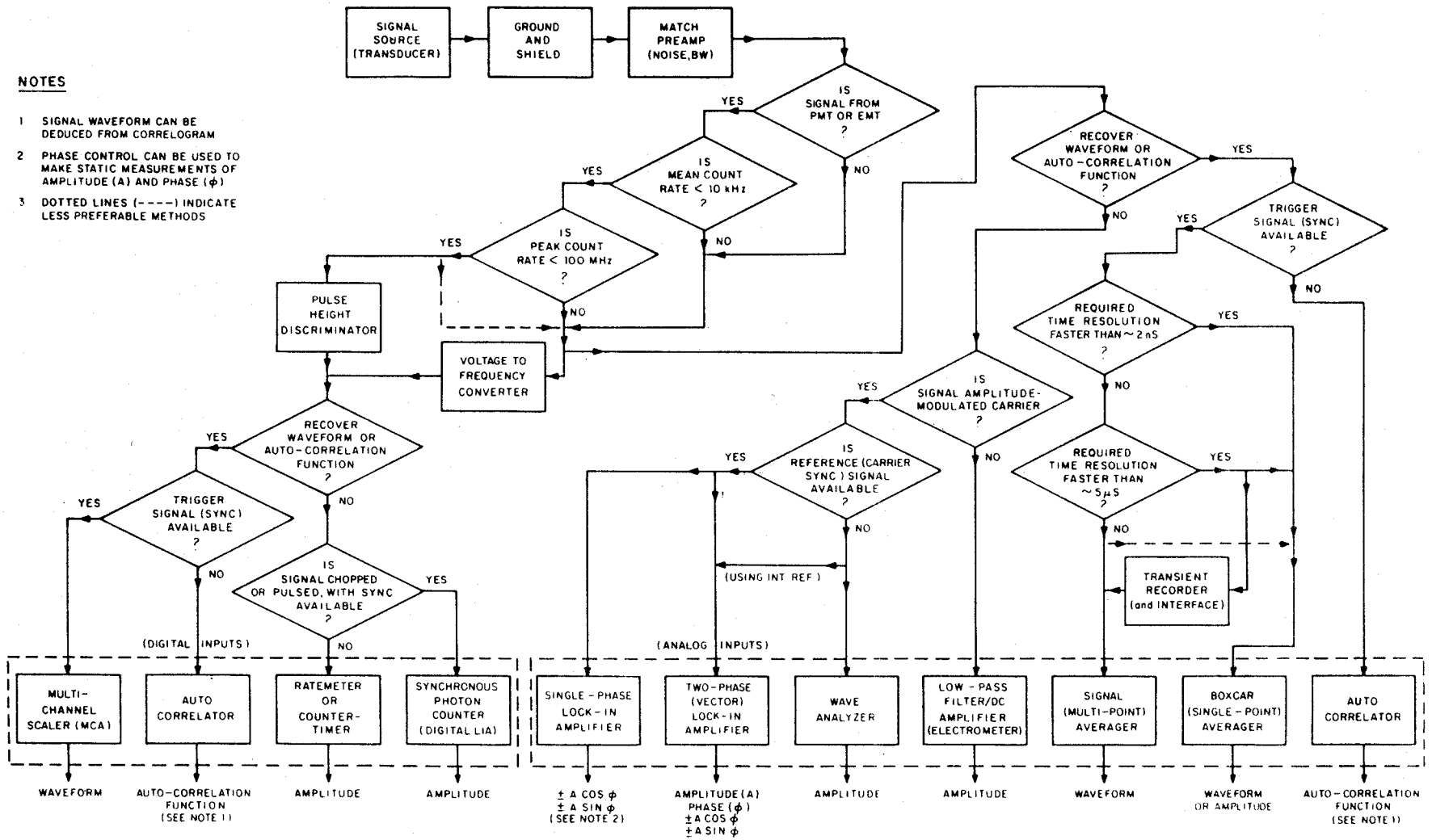



Figure 11.43 Signal recovery flowchart (reproduced by permission of EG&G Princeton Applied Research Corporation)

voltage-to-frequency converter (VFC) ahead of an MCA in MCS mode, analog signals can be time-averaged, and the VFC/MCS combination is essentially a multipoint averaging system. The MCA can also be used in a pulse-height analysis mode, where the amplitude of each input pulse is digitized and used to self-address the memory. In other words, each input pulse with an amplitude between 63.85% and 63.95% (say) of full-scale, will add one count to memory address No. 639. In this way, a pulse-height distribution, or spectrum, is built up. Another common MCA measurement technique is to precede the MCA by a time-to-amplitude converter (TAC), so that each input pulse to be digitized corresponds to a time interval. Low-level measurements of short fluorescent lifetimes, for example, may be made in this fashion.

The photon correlator (Cummins and Pike, 1974) is similar in many ways to the analog-input autocorrelator described in Section 11.9. The input signal is in the form of pulses, from an amplifier–discriminator, and data processing is in serial, rather than parallel, form—with counters replacing the analog correlator's memory. Such digital correlators usually provide at least one mode of *clipped* operation where, for example, $n$ or more input pulses in a lag time ($\tau$) may correspond to a one, and less than $n$ pulses correspond to a zero. Such clipped operation can allow very fast binary shifting and multiplication.

The flowchart of Figure 11.43 makes no attempt to include all instruments or systems. A lock-in amplifier is often used in front of a multipoint averager in order to reduce $1/f$ noise problems. Similarly, a boxcar/multipoint averager combination can offer the picosecond or nanosecond time resolution of the boxcar—without the need to scan so slowly that the system being measured may change during a scan (sweep).

A last comment: the object in signal recovery is not to maximize the SNIR but to minimize the measurement time required to reach a particular output SNR. Similarly, in selecting a preamplifier, the real object is to minimize noise, not noise figure. The noise and/or bandwidth of the signal source or transducer should, therefore, be minimized before seeking instrumentational means of further SNR improvement.

## ACKNOWLEDGEMENTS

The author would like to thank Eric Faulkner, Hans Jorgensen and many other colleagues at EG & G Princeton Applied Research Corporation for numerous discussions and suggestions during the preparation of this manuscript.

## REFERENCES

Blair, D. P. and Sydenham, P. H. (1975). 'Phase sensitive detection as a means to recover signals buried in noise', *J. Phys. E.: Sci. Instrum.*, **8**, 621–7.

Childers, D. G. and Durling, A. E. (1975). *Digital Filtering and Signal Processing*, West Publishing Co., New York.

Cummins, H. Z. and Pike, E. R. (Eds.) (1974). *Photon Correlation and Light Beating Spectroscopy*, Plenum Press, New York.

Darland, E. J., Leroi, G. E., and Enke, C. G. (1979). 'Pulse (photon) counting: determination of optimum measurement systems parameters', *Anal. Chem.*, **51**, 240–5.

EG&G 1. *Operating Manual for Model* 124*A Lock-in Amplifier*, EG&G Princeton Applied Research Corp., Princeton, US.

EG&G 2. *Operating Manual for Model* 162 *Boxcar Integrator*, EG&G Princeton Applied Research Corp., Princeton, US.

EG&G 3. *Operating Manual for Model* 4203 *Signal Averager*, EG&G Princeton Applied Research Corp., Princeton, US.

Faulkner, E. A. (1966). 'Optimum design of low-noise amplifiers', *Electron. Lett.*, **2**, 426–7.

Fellgett, P. B. and Usher, M. J. (1980). 'Fluctuation phenomena in instrument science', *J. Phys. E.: Sci. Instrum.*, **13**, 1041–6.

Hewlett-Packard Co. (1968). *Hewlett-Packard J.*, **19**, 1–16 (whole issue: articles on signal averaging).

Honeywell. 'Correlation and probability analysis', *Saicor Bulletin TB14*, Honeywell Test Instruments Div., US.

Horlick, G. (1975). 'Reduction of quantization effects by time averaging with added random noise', *Anal. Chem.*, **47**, 352–4.

Morrison, R. (1977). *Gounding and Shielding Techniques in Instrumentation*, Wiley, Chichester.

Nicholson, P. W. (1974). *Nuclear Electronics*, Wiley, New York. Chap. 4.