CHAPTER 2: DETECTION

All electromagnetic receiving systems involve *detection*, which is the conversion of a received electrical signal into another electrical signal for interpretation. The second signal typically characterizes the input signal power or the degree of correlation between the received signal and a reference wave form. Detection systems typically consist of one or more physical devices that amplify or convert the signals, unavoidably introducing noise, and other elements that manipulate the signals to achieve the desired output.

This chapter begins in Section 2.1 with a discussion of the major noise processes that limit detector performance. Section 2.2 then analyzes various systems for measuring power or correlations between signals. Section 2.3 characterizes how power and noise propagate within receiver subsystems to define total system performance. Section 2.4 then extends the discussion to detection of optical and infrared signals, for which photon statistics produce different performance characteristics. Physical characteristics of common detectors are also analyzed in an introductory way.

2.1 NOISE PROCESSES

2.1.1. Thermal noise in single-mode transmission lines

In most electromagnetic systems noise limits the performance. This noise has two primary physical origins. First the random thermal motion of charge carriers in these electromagnetic systems induces random electric fields and voltages. Secondly, the natural fluctuations in arrival times of statistically independent quantized charges, photons, or phonons produce "*shot noise*".

The noise associated with radiation from thermally excited charge carriers in solids, liquids, gases, or plasmas is called *thermal noise*, or in the case of random voltage fluctuations in a resistor, *Johnson noise*. The algebraic expressions for the noise depend on the dimensionality of the structure in which it exists. First we shall derive the thermal noise radiated by a simple transmission line, and then thermal radiation in free space and in multimode waveguides. Such radiation propagating through thermally inhomogeneous media is characterized by the equation of radiative transfer. These expressions all take simple forms in their low-frequency and high-frequency limits, and a somewhat more complex form for intermediate infrared wavelengths.

Perhaps the simplest case to understand is that of thermal noise in a one-dimensional transmission line that propagates only one mode, typically the TEM or (Transverse Electromagnetic Mode). It is useful to understand the derivation of the *TEM noise equation* because the same concepts emerge later in other contexts. Our objective is to compute the power

spectral density of the TEM wave emerging from the left-hand end of the transmission line illustrated in Figure 2.1-1.

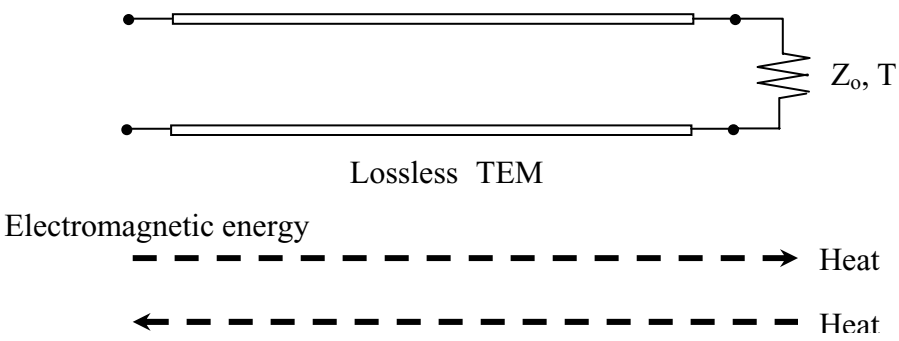

Figure 2.1-1 Coupling between electromagnetic power and heat in a TEM transmission line.

Electromagnetic energy propagated to the right in this matched transmission line will be absorbed by the matched load Z_o and converted to heat. Conversely, thermal motion of charged carriers in the resistor will produce a fluctuating voltage across its terminals which is then perfectly matched to a TEM wave propagating power to the left.

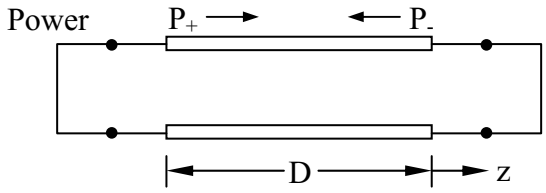

Figure 2.1-2 Closed slightly lossy TEM transmission line resonator.

Our approach to computing the power spectral density propagated toward an observer will be to consider a closed container in thermal equilibrium at temperature T(K) which is very slightly coupled electromagnetically to electromagnetic waves inside; that is, it is very slightly lossy. Figure 2.1-2 illustrates this lossy container of length D. This container exhibits an infinite number of resonant modes and frequencies. To find the average thermal power spectral density P propagating to the right and left, we shall first find the average energy density W(f) [$Jm^{-1}Hz^{-1}$], and then relate this energy density to the average power P_+[W/Hz] propagating to the right.

The time average energy density W(f) [$Jm^{-1}Hz^{-1}$] is readily found:

$$W(f) = \left(\frac{\text{modes}}{\text{Hz}}\right)\left(\frac{\text{photons}}{\text{mode}}\right)\left(\frac{\text{energy}}{\text{photon}}\right) \bullet \frac{1}{D} \tag{2.1.1}$$

where the *energy per photon* is hf [Joules] where *Planck's constant* $h = 6.624\times10^{-34}\,(\mathrm{Js})$ and f is the photon frequency (Hz).

To find the *mode density* or number of modes per Hertz we note that the short circuits at each end of the transmission line force the voltage there to be zero. Therefore at each resonance frequency f_m, there must be an integral number of half wavelengths along the transmission line, as suggested in Figure 2.1-3.

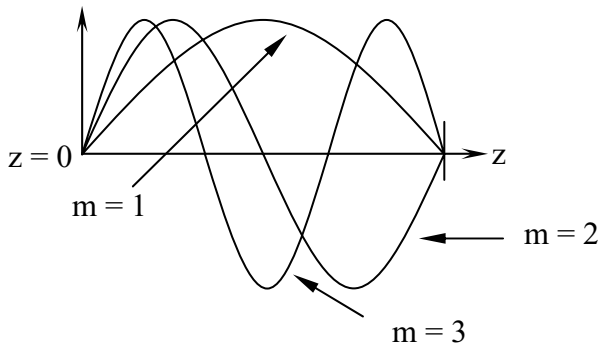

Figure 2.1-3 TEM resonator modes.

The number of half wavelengths in length D is

$$m = \frac{2D}{\lambda_m} = \frac{2Df_m}{v_p} \tag{2.1.2}$$

where v_p is phase velocity. The number of modes per Hertz is readily found by differentiating Equation (2.1.2) with respect to frequency:

$$\frac{dm}{df} = \frac{2D}{v_p}\left(\mathrm{modes/Hz}\right) \tag{2.1.3}$$

The average number of photons $\bar{n}_j$ in the jth mode can be computed if we know the probability $p_j(n)$ of having n photons in mode j.

Because photons obey *Bose-Einstein statistics*, any number of photons can occupy each mode. Since the total energy in the system is fixed, and since combinatorics favor the more likely distributions, the probability distribution of photons among states is proportional to $e^{-nW_j/kT}$, which is called the *Boltzmann distribution*, where the *Boltzmann constant* $k = 1.3805\times10^{-23}\,(\mathrm{J/K})$, $W_j = hf_j$, and the total energy in state j is nW_j. We can thus express the probability distribution of photons among states in terms of a proportionality constant Q that must be determined:

$$p_j(n) = Qe^{-nW_j/kT} = Q \bullet \sum_{n=0}^{\infty} \left(e^{-W_j/kT}\right)^n = \frac{Q}{1-e^{-W_j/kT}} \tag{2.1.4}$$

where

$$\sum_{n=0}^{\infty} p_j(n) \equiv 1 \tag{2.1.5}$$

$$\sum_{n=0}^{\infty} x^n = 1/(1-x) \quad \text{if } x<1 \tag{2.1.6}$$

therefore

$$Q = 1 - e^{-W_j/kT} \tag{2.1.7}$$

$$p_j(n) = \left(1-e^{-W_j/kT}\right)e^{-nW_j/kT} \tag{2.1.8}$$

$$\bar{n}_j = \sum_{n=0}^{\infty} n\, p_j(n) = \left(1-e^{-W_j/kT}\right)\sum_{n=0}^{\infty} n\left(e^{-W_j/kT}\right)^n \tag{2.1.9}$$

The sum of (2.1.9) can be evaluated by recalling:

$$\sum_{n=0}^{\infty} n\, x^n = x\frac{d}{dx}\sum_{n=0}^{\infty} x^n = x\frac{d}{dx}(1-x)^{-1} = \frac{x}{(1-x)^2} \tag{2.1.10}$$

Thus

$$\bar{n}_j = \left(1-e^{W_j/kT}\right)\left[e^{-W_j/kT}\Big/\left(1-e^{-W_j/kT}\right)^2\right] \tag{2.1.11}$$

The average number of photons $\bar{n}_j$ in state j, or *photon state density*, can be more simply written as

$$\bar{n}_j = 1\Big/\left(e^{hf_j/kT}-1\right) \tag{2.1.12}$$

where hf_j is the energy W_j associated with the state j.

We now have all the elements of Equation (2.1.1) for the average energy density W(f) [J/m Hz] in the closed TEM resonator, which is:

$$W(f) = \left(\frac{2D}{v_p}\right)\left(\frac{1}{e^{W_j/kT} - 1}\right)(hf) \bullet \frac{1}{D} = \frac{2hf}{v_p\left(e^{hf/kT} - 1\right)} \ [Jm^{-1}Hz^{-1}] \tag{2.1.13}$$

Now we can relate this energy density to the power flow inside the closed resonator. The total energy density W(f) can be associated with uncoupled energy flows to the right and left characterized by $W_+ + W_- = W(f) = 2W_+[J/m]$. The power flowing to the right $P_+[W/Hz]$ is simply the group velocity v_g times W_+. If the transmission line is nondispersive then the group velocity v_g equals the phase velocity v_p in Equation (2.1.13), which leads to:

$$P_+(f)\left[WH_z^{-1}\right] = \frac{hf}{e^{hf/kT} - 1} \tag{2.1.14}$$

This simple expression for *thermal noise per mode* is characterized by Figure 2.1-4

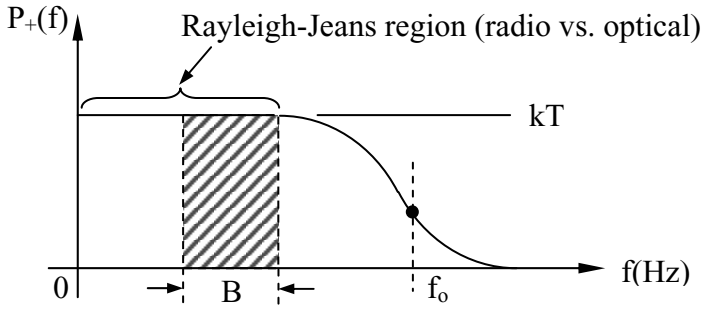

Figure 2.1-4 Thermal power spectral density $P_+[W/Hz]$ in a TEM transmission line.

The figure suggests there is a white-noise region for frequencies f where $hf << kT_o$. This inequality is called the *Rayleigh-Jeans approximation* and applies to the low frequency end of the spectrum, or typically the radio spectrum as opposed to the optical spectrum. The transition frequency f_o between these two regions is temperature dependent, where:

$$f_o = kT/h \tag{2.1.15}$$

Equation (2.1.15) says that the transition frequency f_o(GHz) is approximately 20 times the temperature T of the object in degrees Kelvin. This transition frequency at normal 300K temperatures is approximately 6 THz, in the infrared region.

A simple expression for the thermal power spectral density in the Rayleigh-Jeans limit can be found by replacing the exponential in Equation (2.1.14) using the series expansion $e^x = 1 + x + x^2/2! + ... = \sim 1 + x$ for $x << 1$. This results in the expression for power spectral density in the Rayleigh-Jeans limit:

$$P_+(f)\left[\mathrm{WHz}^{-1}\right] \cong kT \text{ for } hf << kT \quad \text{"Rayleigh-Jeans limit"} \tag{2.1.16}$$

In the Rayleigh-Jeans limit it becomes trivial to compute the total thermal power within some bandwidth B[Hz]. In this limit the *thermal power* propagating down a single-mode transmission line from a matched load at temperature T is:

$$P \cong kTB \text{ watts} \tag{2.1.17}$$

This equation is widely used in characterizing radio systems.

2.1.2. Thermal radiation in space

The intensity of thermal radiation propagating in three-dimensions in free space can be found using a similar derivation but beginning with a three-dimensional lossless resonator in thermal equilibrium at temperature T. In this case we relate the energy density spectrum W(f) [$\mathrm{Jm^{-3}Hz^{-1}}$] to the *thermal radiation intensity* I [$\mathrm{Wm^{-2}Hz^{-1}ster^{-1}}$].

To find the energy density spectrum W(f) we modify (2.1.1) by dividing by resonator volume instead of by the transmission line length:

$$W(f) = \left(\frac{\text{modes}}{\text{Hz}}\right) \bullet \left(\frac{\text{photons}}{\text{mode}}\right) \bullet \left(\frac{\text{energy}}{\text{photon}}\right) \bullet \frac{1}{\text{vol.}} \tag{2.1.18}$$

This time our resonator is a slightly lossy rectangular conducting box of dimensions $a \times b \times d$. Resonances in such a box have integral numbers of half wavelengths in each of the three dimensions, where we assume m, n, and p half wavelengths are associated with the dimensions a, b, and d, respectively. Thus $a = m\lambda_y/2$, $b = n\lambda_x/2$, and $d = p\lambda_z/2$. The resonant frequency (Hz) for the mode m, n, p is then:

$$f = \sqrt{\left(\frac{mc}{2a}\right)^2 + \left(\frac{nc}{2b}\right)^2 + \left(\frac{pc}{2d}\right)^2} \tag{2.1.19}$$

This expression for resonant frequency is readily derived by substituting the expression for a uniform planewave,

$$\underline{\overline{E}} = \underline{\overline{E}}_o e^{-jk_x x - jk_y y - jk_z z} \tag{2.1.20}$$

into the wave equation:

$$\begin{aligned} &\left(\nabla^2 + \omega^2 \mu\varepsilon\right)\underline{\overline{E}} = 0 \\ &\downarrow \\ &\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) \end{aligned} \tag{2.1.21}$$

The operator ∇^2 sums the second spatial derivatives with respect to each of x, y, and z, yielding:

$$k_x^2 + k_y^2 + k_z^2 = k_o^2 = \omega^2 \mu_o \varepsilon_o = (2\pi f/c)^2 \tag{2.1.22}$$

Since $k_x = 2\pi/\lambda_x$, where $\lambda_x = 2b/n$, Equation (2.1.19) follows directly from (2.1.22).

We can represent (2.1.19) graphically as shown in Figure 2.1-5

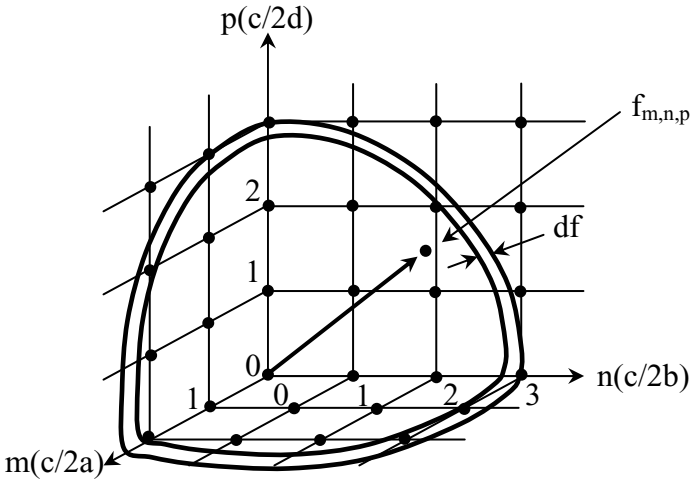

Figure 2.1-5 Spectral density of resonant frequencies of a cavity.

The frequency $f_{m,n,p}$ is equivalent to the radial distance of the m, n, p node from the origin in the figure, where each node corresponds to one combination of the quantum numbers m, n, and p. Each such node corresponds to both a TE and a TM resonance, where we may think of the rectangular cavity as a waveguide propagating the $TE_{m,n}$ and $TM_{m,n}$ waveguides modes inside a waveguide of dimensions $a \times b$. The dimensions of each unit cell in the figure are c/2a, c/2b, and c/2d, as suggested by (2.1.19).

If the mode numbers are high, then to a good approximation the number of resonant modes falling within a frequency interval df is simply the volume of a thin spherical shell of thickness df, as suggested in Figure 2.1-5. The number of modes is that volume divided by the volume per unit cell, multiplied by two to account for both the TE and TM modes associated with each combination m, n, p. That is, the number of modes in such a shell are:

$$\frac{4\pi f^2 df \bullet 2}{8} \Bigg/ \left(\frac{c}{2a} \bullet \frac{c}{2b} \bullet \frac{c}{2d} \right) = \frac{8\pi f^2}{c^3} V\, df \Rightarrow \left[\frac{\text{modes}}{\text{Hz}} \right] df \tag{2.1.23}$$

where the volume V of the resonator is abd.

Substituting (2.1.23) into (2.1.18) yields the expression for the *thermal energy density spectrum* W(f):

$$W(f) = \left(\frac{8\pi f^2}{c^3} V \right) \left(\frac{1}{e^{hf/kT} - 1} \right) \bullet hf \bullet 1/V = \frac{8\pi}{c^3} \frac{hf^3}{e^{hf/kT} - 1} \ [\text{Jm}^{-3}\text{Hz}^{-1}] \tag{2.1.24}$$

The energy density spectrum W(f) can now be related to the radiation intensity I(θ,ϕ,f) by imagining a thin slab of unit area and thickness δ which contains Wδ[J/Hz]. If the radiation in the slab is in thermal equilibrium at temperature T, and then radiates away without replacement at all angles θ from the normal, then we may compute Wδ in two ways:

$$W\delta = \int_V W(f)\, dV = \int I(f)\ dA\ dt\ d\Omega \quad [\text{J/Hz}] \tag{2.1.25}$$

At angle θ, the projected area of the slab is cos θ and the intensity is I_o. Because the radiation escapes from the slab without replacement, the pulse in any direction lasts for δ/(c cos θ) seconds. (2.1.25) then becomes:

$$W\delta = \delta \frac{8\pi}{c^3} hf^3 \Big/ \left(e^{hf/kT} - 1 \right) = \int_{4\pi} I_o \cos\theta \left(\delta / c \cos\theta \right) d\Omega \tag{2.1.26}$$

Planck's radiation law for the intensity of blackbody thermal radiation then follows directly from (2.1.26):

$$I_o(f, \theta, \phi) = 2hf^3 \Big/ \left(e^{hf/kT} - 1 \right) \ [\text{Wm}^{-2}\,\text{Hz}^{-1}\,\text{ster}^{-1}] \tag{2.1.27}$$

In the low frequency limit where $hf << kT$, Planck's law for thermal radiation reduces to the *Rayleigh-Jeans law*:

$$I_o(f,\theta,\phi) \cong \frac{2kT}{\lambda^2} \ [Wm^{-2}\,Hz^{-1}\,ster^{-1}] \tag{2.1.28}$$

The frequency dependence of Planck's law and its Rayleigh-Jeans approximation is suggested in Figure 2.1-6.

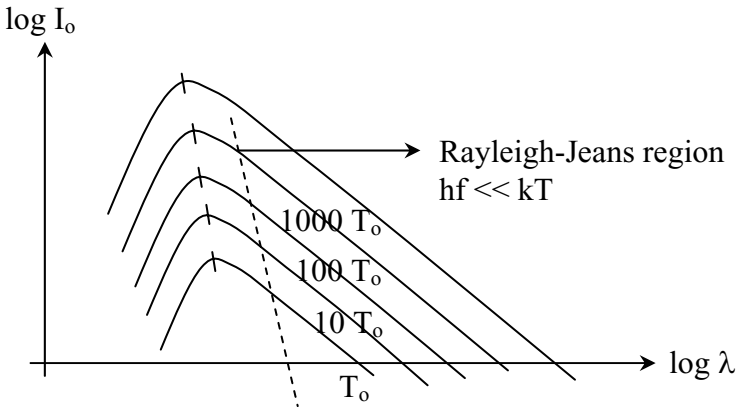

Figure 2.1-6 Wavelength dependence of Planck's radiation law.

Planck's law peaks in the microwave region for cryogenic temperatures, and moves into the infrared band for normal environmental temperatures. The peak shifts into the visible band at furnace temperatures, moving from redhot to whitehot as temperatures climb towards solar values of thousands of degrees. In the Rayleigh-Jeans region the intensity ($Wm^{-2}\,Hz^{-1}\,ster^{-1}$) is directly proportional to kinetic temperature and inversely proportional to the square of the wavelength. This thermally linear region is evident in Figure 2.1-6.

We might suppose there is a paradox in the frequency independence of thermal radiation found in (2.1.16) and the dependence found in (2.1.28), as illustrated in Figure 2.1-7 which shows a transmission line coupled to free space.

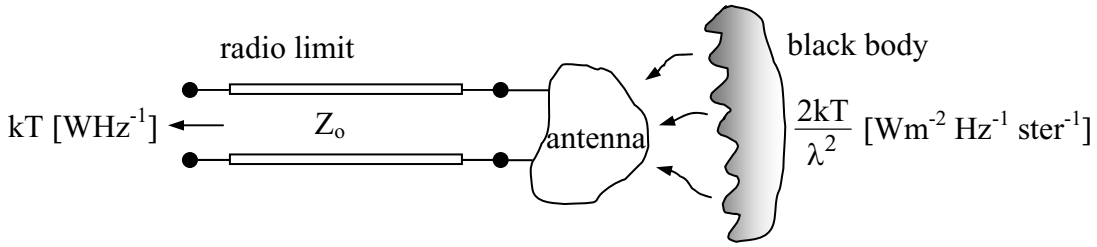

Figure 2.1-7 Wavelength-dependence paradox for thermal radiation.

The resolution of the paradox is simple because in the Rayleigh-Jeans limit the intensity is kT [W/Hz] per mode in both transmission lines and free space; free space simply has $2/\lambda^2$ modes m^{-2} $ster^{-1}$, where the factor 2 corresponds to the two possible polarizations, TE and TM.

It is interesting to note that thermal radiation in a TEM cavity possesses kT/2 Joules per degree of freedom, the same equilibrium value exhibited by particles in *Brownian motion*. This can be easily shown by noting the average energy per mode is simply proportional to the average number of photons in a mode j, and equal to:

$$hf_j\bar{n}_j = hf\frac{1}{e^{hf/kT}} \cong kT \ \left[J/\text{mode}\right] \tag{2.1.29}$$

Since each mode has two degrees of freedom, proportional to sin ωt and cos ωt, each degree of freedom has average energy kT/2.

2.1.3. Thermal noise in circuits

Thermal noise can originate from resistors as well as from transmission lines and free space. For example, consider the resistor R matched to a transmission line as illustrated in Figure 2.1-8.

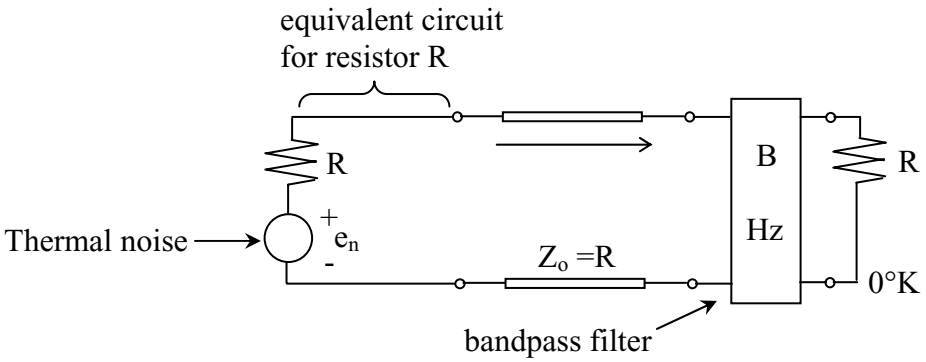

Figure 2.1-8 Thermal noise voltages (Johnson noise) produced by resistors.

The resistor R can be replaced by a passive resistor R and a thermal noise voltage source producing an rms noise voltage e_n. This resistor may then be coupled to a bandpass filter of bandwidth B Hertz by a lossless matched transmission line having impedance Z_o = R. The thermal noise power reaching resistor R at the righthand side at 0 K is kTB watts, which equals $(e_n/2)^2/R$ because within the bandwidth B the thermal noise voltage is divided across the two resistors R in series. Solving for the rms thermal noise voltage e_n we find:

$$e_n\ (\text{thermal noise}) = \sqrt{4\,kTBR}\ \text{volts (in B Hz)} \tag{2.1.30}$$

The expression for thermal noise, also known as *Johnson noise*, in (2.1.30) can, for example, easily yield the thermal noise at the input of a 50-ohm input amplifier with a bandwidth of 100MHz and an equivalent temperature of 300K:

$$e_n = \sqrt{4 \times 1.38 \times 10^3 \times 300 \times 10^8 \times 50} = 9.1\,\mu v \tag{2.1.31}$$

Thermal voltages can become quite substantial across large resistors; for example, (2.1.31) yields 9.1 mv for R = 50 M ohm.

Lossy media such as transmission lines can also radiate, as suggested in Figure 2.1-9.

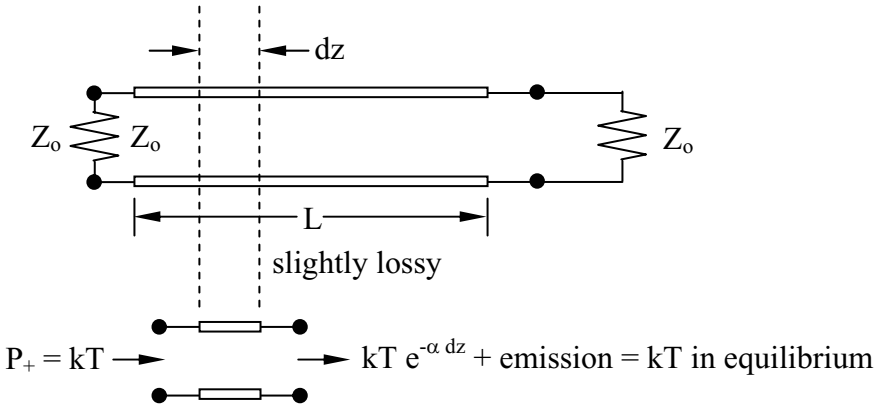

Figure 2.1-9 Thermal radiation emitted by lossy TEM transmission lines.

We assume a slightly lossy TEM transmission line is terminated at both ends with matched loads of impedance Z_o, and all are in equilibrium at temperature T K. In equilibrium the power P_+ moving to the right must everywhere equal kT, even though a short lossy segment of length dz absorbs some of that power, and propagates to the right $kTe^{-\alpha dz}$. To maintain thermal equilibrium the short segment dz must therefore emit sufficient radiation to maintain the equilibrium kT. Since αdz is quite small, we can truncate the power series expansion for the exponential to find the emission term:

$$\text{Emission } [W/Hz] = kT\left(1\text{-}e^{-\alpha dz}\right) \cong kT\left(1-[1-\alpha dz]\right) = kT\alpha\, dz \tag{2.1.32}$$

Since the emission by the thermal line does not depend on the intensity of the radiation passing through it, we can use (2.1.32) to calculate the thermal emission emerging from a transmission line which has a nonuniform temperature distribution. The output thermal power spectral density kT_{out} equals the attenuated input emission, plus emission contributed by each incremental segment of the line, as characterized by (2.1.32), and attenuated by the length of line between its source and the output. Thus:

$$kT_{out} = kT_{in}e^{-\int_o^L \alpha dz} + k\int_o^L T(z)\alpha(z)e^{-\int_z^L \alpha dz}dz \tag{2.1.33}$$

This expression can be simplified by recasting it in terms of an equivalent temperature T_{out} and *optical depth* $\tau = \int_o^z \alpha dz$:

$$T_{out} = T_{in}e^{-\int_o^L \alpha dz} + \int_{\tau_{max}}^{o} T(\tau)e^{-\tau}d\tau \tag{2.1.34}$$

where $\tau_{max} = \int_o^L \alpha dz$. Equation 2.1.34 is the *equation of radiative transfer* and applies not only to TEM transmission lines, but also to propagation of waves through free space and other transmission media. A simplification which applies to transmission lines of constant temperature and optical depth τ is:

$$T_{out} = T_{in}e^{-\tau} + T_{line}\left(1 - e^{-\tau}\right) \tag{2.1.35}$$

Consider the simple example illustrated in Figure 2.1-10 of a receiver connected to an antenna viewing cold space, for which the equivalent thermal temperature (brightness temperature) is 3K.

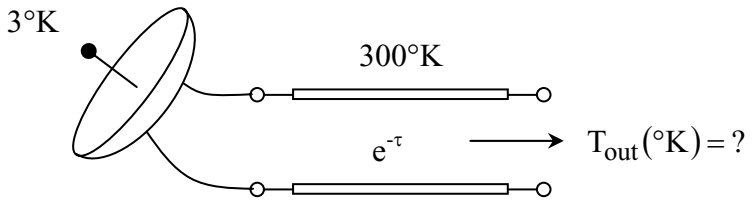

Figure 2.1-10 Effect of a lossy transmission line on antenna signals.

If the optical depth τ of the transmission line is zero, then (2.1.35) suggests the output temperature T_{out} would be 3K in this case. If $\tau = \infty$, then $T_{out} = 300K$. If the transmission line exhibits 2 dB loss, then $e^{-\tau} = 10^{-2/10} = 0.63$, so that $T_{out} = 3 \times 0.63 + 300\left(1 - 0.63\right) \cong 113K$.

2.1.4. Shot noise

Perhaps the second most important type of noise in addition to thermal noise is called *shot noise*, because it sounds like falling shot. One way to produce shot noise is illustrated in Figure 2.1-11 for a *vacuum tube diode*.

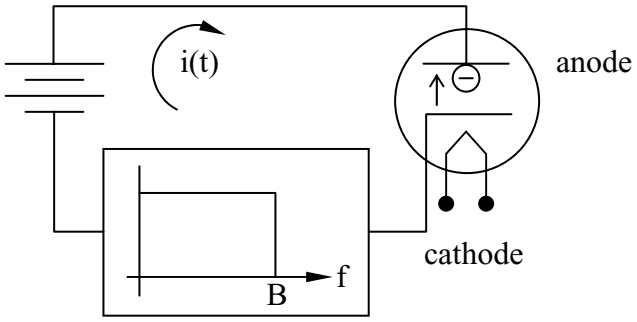

Figure 2.1-11 Shot noise production by discrete electron conduction in a vacuum tube.

The current flow through this or other types of diode consists of multiple current pulses, the integral of each being one electron charge. Usually these electrons move across the diode independently in Poisson-distributed fashion. The average current is the charge on an electron times the average number of electrons passing per second. The following derivation of the shot-noise power density spectrum requires these electron transits to be independent, which excludes high-current vacuum tubes where the current flow is smoothed by electrons piling up in the transit path so as to modulate the electric fields seen by the individual charge carriers. Under the assumption of independent electron arrival times it is a straightforward matter to calculate the power spectral density $\Phi_i(f)$ of the current i(t) as the Fourier transform of the current's autocorrelation function $\phi_i(\tau)$.

Although it can readily be shown that the following result is true for any shape of current pulse i(t) associated with a single electron, the derivation is trivial if we assume it has a boxcar form, as illustrated in Figure 2.1-12.

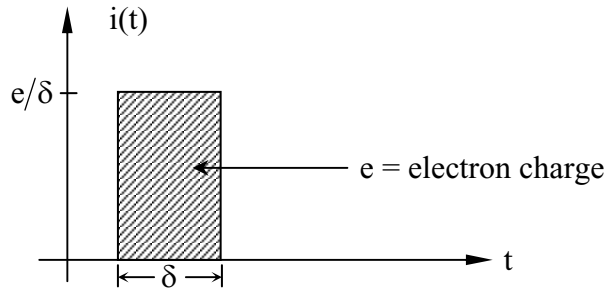

Figure 2.1-12 Idealized current pulse shape for calculating shot noise in diodes.

The autocorrelation function $\phi_i(\tau)$ can be found as suggested in Figure 2.1-13.

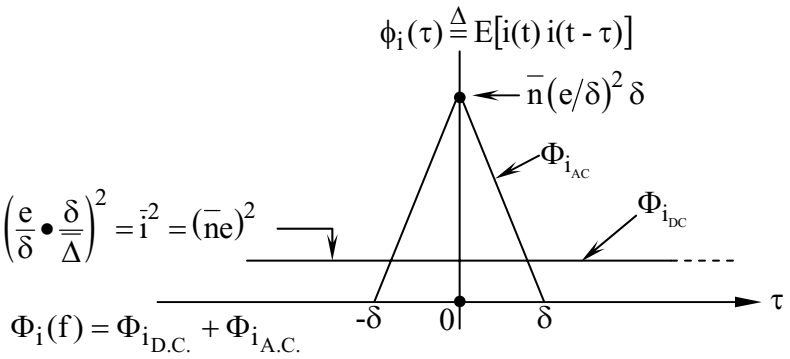

Figure 2.1-13 Autocorrelation function $\phi_i(\tau)$ of a Poisson-distributed series of boxcar current pulses corresponding to independent electron transits.

The constant term in the autocorrelation function is associated with the possibility that the current pulse e/δ from one electron will overlap a current pulse from another electron, the probability of which is $\delta/\bar{\Delta}$, where $\bar{\Delta}$ is defined as the average time between current pulses, i.e. $\bar{\Delta} = 1/\bar{n}$.

The power spectral density $\Phi_i(f)$ has both a DC part an AC part, as shown in Figure 2.1-14.

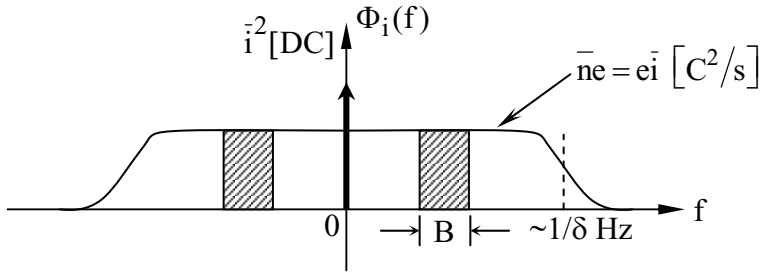

Figure 2.1-14 Shot-noise power spectral density for an average current

If we are interested in signal components below frequencies of $\sim 1/\delta$ Hz, then we will incur additive shot noise associated with $\Phi_i(f)$ within the bandwidth B. The white-noise portion of the shot noise equals the integral under the triangular impulse illustrated in Figure 2.1-13, which is $\bar{n}e^2 = e\bar{i}\ \left[C^2/s\right]$. Therefore the variance $\sigma^2_{i_{AC}}$ associated with the shot noise within a bandwidth B Hz is the double-sided integral of the power spectral density $e\bar{i}$ over the bandwidth B, yielding the standard *shot noise variance*:

$$\sigma^2_{i_{A.C.}} = 2Be\bar{i}\left[Amp^2\right] \tag{2.1.36}$$

Consider a simple example where the output voltage across a 5-KΩ resistor is being measured within a one-megahertz bandwidth B and where the one-milliampere average current through the resistance R is limited by a diode that passes electrons with Poisson-distributed arrival times. The rms shot noise across the resistor R is:

$$v_{rms}(\text{shot}) = \sqrt{\sigma_{i_{AC}}^2}\, R \cong 0.1\,\text{mv} \tag{2.1.37}$$

This can be compared to the Johnson noise for T = 300K, given by:

$$v_{rms}(\text{thermal}) = \sqrt{4kTBR} \cong 0.01\,\text{mv} \tag{2.1.38}$$

The average voltage across the given output resistor is $\bar{i}R = 10^{-3} \times 5K = 5$ volts, very large compared to the ~0.1-mv Johnson noise, and large compared to the ~0.1-mv shot noise. Depending on the system parameters, either the thermal noise or the shot noise may actually dominate.

Although shot noise can consist of well-separated impulses, more generally the electron arrival rate is so great that their pulses substantially overlap. The voltage at any instant then is the sum of independent random events which, by the central limit theorem, approaches a Gaussian distribution as the number of overlapping events increases. Thus shot noise is typically Gaussian white noise below some frequency and is therefore indistinguishable from Johnson noise in most practical situations.

2.2 POWER SPECTRAL MEASUREMENT OF RADIO SIGNALS

2.2.1 Measurement of thermal power

Measurements of power (by a radiometer) or power spectral density (by a spectrometer) are perhaps the most fundamental of observations. For example, *Morse code* is communicated by sending dots and dashes separated by periods of silence; this is a form of on-off keying. The task of the receiver is to determine whether the power level at any instant corresponds to transmission or silence, where the noise is generally dominated by Johnson or shot noise in the absence of interference. Frequency-shift-keyed transmissions jump from frequency to frequency, typically conveying binary information to receivers observing the power in each of two or more frequency channels perturbed by additive Gaussian noise. Radio astronomers and other observers of the physical environment or telecommunications activity frequently wish to measure power spectral densities, and may use a variety of spectrometers for this purpose. The design and performance of such systems is the subject of the following section.

One standard objective is measurement of the power in an incoming signal in a defined bandwidth of B Hz, averaged over a period of τ seconds. A standard system for making such measurements is called a *total power radiometer*. It simply computes the average value of the square of a voltage after it passes through a bandpass filter of bandwidth B, as suggested in Figure 2.2-1. Thus the total power radiometer simply computes the definition of average power over some time interval τ which characterizes the impulse response h(t) of the output filter (typically a boxcar filter of duration τ).

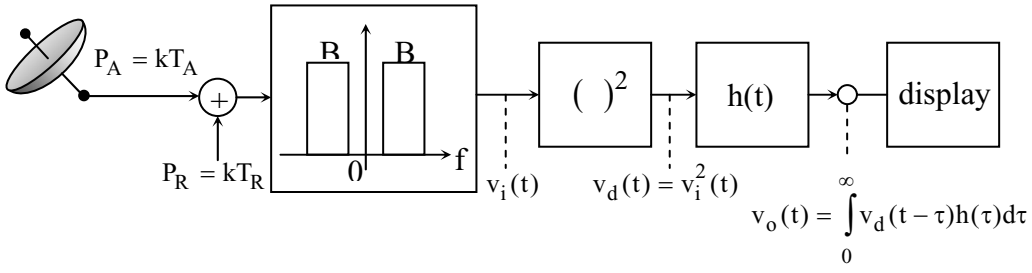

Figure 2.2-1 Total power radiometer block diagram.

Assume the power spectral density P_A entering the antenna is simply kT_A watts per Hertz, where T_A is the *antenna temperature* as seen from the antenna output and is the parameter we wish to measure. The Gaussian white noise P_R added by the radiometer due to thermal and shot noise processes is characterized by kT_R. The sum $k(T_A+T_R)$ then passes through the bandpass filter of width B Hz. The voltage waveform entering the filter is suggested in the upper left hand corner of Figure 2.2-2, while $v_i(t)$ emerging from the filter is suggested by the more nearly monochromatic waveform illustrated in Figure 2.2-2. The detected voltage $v_d(t)$ emerging from the square law device, and the output voltage $v_o(t)$ representing the estimated received power emerging from the lowpass filter characterized by its impulse response h(t), are also shown in Figure 2.2-2. Because the output voltage is the smoothed estimate of the square of the bandpass signal, it is proportional to $T_A + T_R$. The lowpass filter averages many independent cycles of the detected voltage and becomes Gaussian by virtue of the central limit theorem, as illustrated in the bottom graph of Figure 2.2-2.

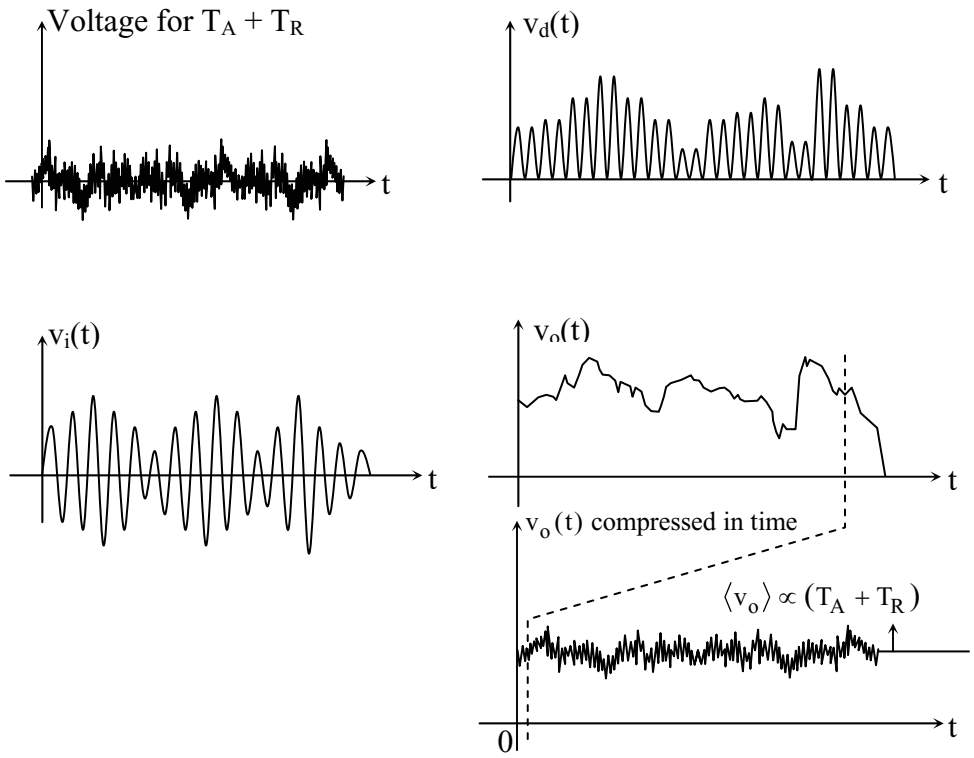

Figure 2.2-2 Signal processing progression in a total power radiometer.

We may calculate the sensitivity of the total power radiometer by evaluating the rms output voltage $v_{o_{rms}}$ and then expressing it terms of equivalent degrees Kelvin, where the output voltage has been calibrated in terms of the desired antenna temperature, or perhaps in some other units. If antenna temperature is the desired output parameter, then the receiver sensitivity is generally expressed in terms of ΔT_{rms}, where:

$$\Delta T_{rms} \triangleq \frac{v_{o_{rms}}}{\frac{\partial \langle v_o \rangle}{\partial T_A}} \tag{2.2.1}$$

where $\partial \langle v_o \rangle / \partial T_A$ calibrates voltage fluctuation in terms of temperature. To evaluate (2.2.1) we first need to calculate $\langle v_o \rangle$ and $v_{o_{rms}}$, where both can be found from the power spectral density $\Phi_o(f)$ of the output signal $v_o(t)$. The desired $v_{o_{rms}}$ can be found from the AC portion of the power spectral density $\Phi_o(f)$ of the output voltage, while $\langle v_o \rangle$ is the DC component of $\Phi_o(f)$. $\Phi_o(f)$ can readily be found, as explained below, from the power spectral density of the detector output, which is the Fourier transform of the detector autocorrelation function.

To find $\Phi_i(f)$ we begin by finding the detector voltage $v_d(t)$ and its autocorrelation function $\phi_d(\tau)$, where:

$$\phi_d(\tau) = E\left[v_d(t) \bullet v_d(t-\tau)\right] = E\left[v_i^2(t)\, v_i^2(t-\tau)\right] \tag{2.2.2}$$

Since $v_d(t)$ is not Gaussian, computing its autocorrelation function is difficult. Fortunately $v_i(t)$ and $v_i(t - \tau)$ are jointly Gaussian random variables with zero mean (abbreviated JGRVZM). In this special case we can compute the expected value of the product of four such JGRVZM using:

$$E[wxyz] = E[wx]E[yz] + E[wy]E[xz] + E[wz]\ E[xy] \tag{2.2.3}$$

It easily follows from (2.2.2) and (2.2.3) that:

$$\phi_d(\tau) = \overline{v_i^2(t)}\,\overline{v_i^2(t-\tau)} + 2\ \ \overline{v_i(t)\, v_i(t-\tau)}^2 = \phi_i^2(0) + 2\phi_i^2(\tau) \tag{2.2.4}$$

where the overbar is an abbreviation for expected value, and we are taking advantage of the fact that the noise signals we are analyzing here are ergodic, which means that time averages of the products of two signal samples (e.g. sampled at times t and t-τ) equal the ensemble average of the same product. Since the Fourier transform of the product of two time functions equals the convolution of their transforms, we can readily compute the Fourier transform of (2.2.4):

$$\Phi_d(f) = \phi_i^2(0)\ u_o(f) + 2\Phi_i(f) * \Phi_i^*(f) \tag{2.2.5}$$

To compute $\Phi_d(f)$ we first must find $\phi_i(f)$, where:

$$\phi_i(0) = \overline{v_i^2(t)} = \int_{-\infty}^{\infty} \Phi_i(f)\, df = kT_{eff}B \tag{2.2.6}$$

where the *effective system temperature* T_{eff} is defined as $T_A + T_R$, and where the power spectral density of the input signal $\Phi_d(f)$ is shaped by the assumed boxcar bandpass filter, as illustrated in Figure 2.2-3.

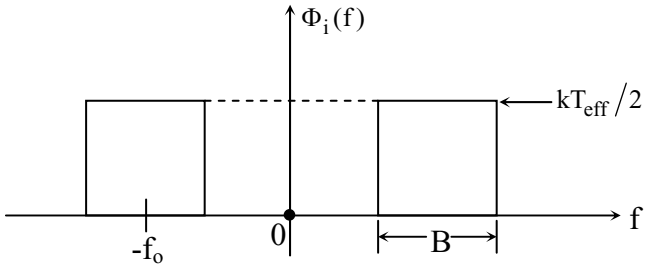

Figure 2.2-3 Power spectral density of a total power radiometer with bandwidth B Hz.

Equation (2.2.5) for $\Phi_d(f)$ includes an impulse plus the convolution of $\Phi_i(f)$ with itself, which can easily be found from (2.2.6) and examination of Figure 2.2-3, as illustrated in Figure 2.2-4. The output signal $v_o(t)$ is the convolution of the detected signal $v_d(t)$ with the impulse response h(t) of the output filter of the total power radiometer:

$$v_o(t) = v_d(t) * h(t) \tag{2.2.7}$$

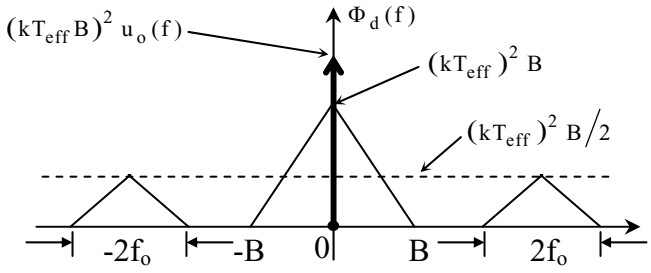

Figure 2.2-4 Power spectral density of the detected signal in a total power radiometer.

Therefore the output power spectral density is:

$$\Phi_o(f) = \Phi_d(f) \bullet |H(f)|^2 \tag{2.2.8}$$

The output filter transfer function $|H(f)|^2$ can be easily found, as suggested in Figure 2.2-5

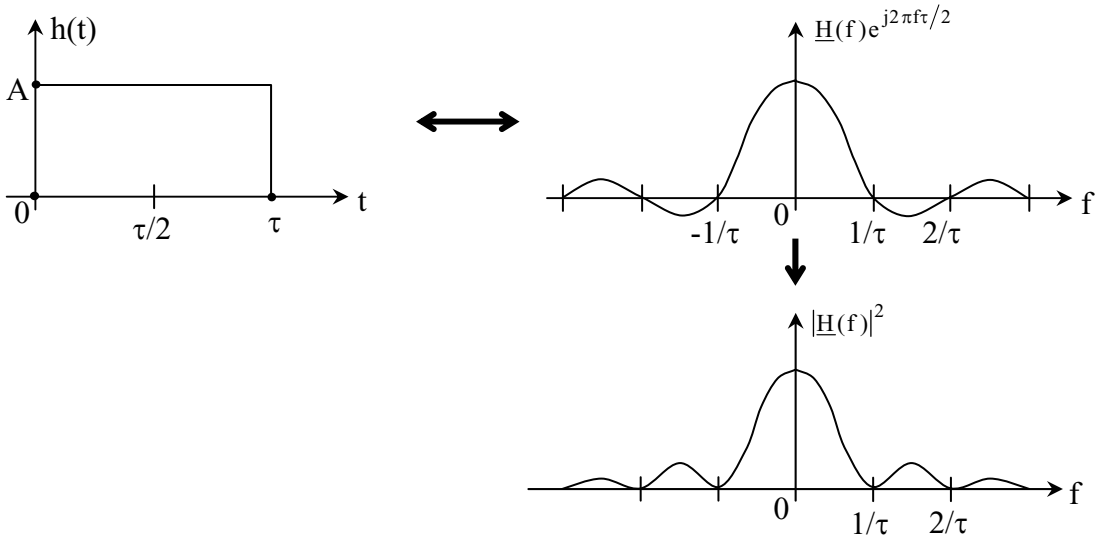

Figure 2.2-5 Integrator impulse response h(t) and its spectral transfer function $|\underline{H}(f)|^2$.

To evaluate (2.2.8) for Figure 2.2-4 we need to know $|\underline{H}(f)|^2$ for f = 0, and the total integral over frequency of $|\underline{H}(f)|^2$. Note that generally B>>1/τ, so that $\Phi_d(f)$ is approximately constant where $|\underline{H}(f)| \neq 0$. The first quantity is easy to find:

$$\underline{H}(f=0) = \int_{-\infty}^{\infty} h(t) e^{-j2\pi(f=0)t} dt = A\tau \tag{2.2.9}$$

Thus the DC power emerging from the output filter is:

$$\Phi_{o_{DC}}(f) = (kT_{eff}B)^2 (A\tau)^2 u_o(f) \tag{2.2.10}$$

The variance of the fluctuating component of the output voltage is:

$$P_{o_{AC}} = \int_{-\infty}^{\infty} \Phi_{o_{AC}}(f) df \cong (kT_{eff})^2 B \bullet \int_{-\infty}^{\infty} |\underline{H}(f)|^2 df \tag{2.2.11}$$

where we have used the fact that for most total-power radiometers the integration time τ is sufficiently large that $1/\tau << B$; therefore only $\Phi_d(f=0)$ is important. By Parseval's theorem:

$$\int_{-\infty}^{\infty} |H(f)|^2 \, df = \int_{-\infty}^{\infty} h^2(t)\, dt = A^2\tau \tag{2.2.12}$$

The desired sensitivity ΔT_{rms} for a total-power radiometer then follows from (2.2.1), (2.2.10), and (2.2.11):

$$\Delta T_{rms} = \frac{\sqrt{P_{AC}}}{\left(\partial\sqrt{P_{DC}}/\partial T_A\right)}\left[^\circ K\right] = \frac{\sqrt{\left(kT_{eff}\right)^2 B \bullet A^2\tau}}{\left(\partial\left[kT_{eff}BA\tau\right]/\partial T_A\right)} = \frac{kT_{eff}A\sqrt{B\tau}}{kA\ B\tau} \tag{2.2.13}$$

Therefore the total-power radiometer sensitivity is:

$$\Delta T_{rms} = \frac{T_A + T_R}{\sqrt{B\tau}} \tag{2.2.14}$$

This expression for sensitivity applies to any receiver employing a square-law detector of signals with additive Gaussian noise.

This expression (2.2.14) can readily be recomputed for other output integrators such as a conventional single-pole RC filter. The impulse response h(t) for such a simple filter is illustrated in Figure 2.2-6.

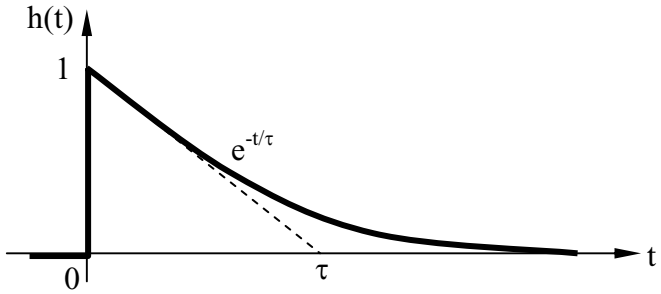

Figure 2.2-6 Impulse response of an RC integrator.

For this RC filter we simply recompute (2.2.8) and note:

$$H(f=0) = \int_{-\infty}^{\infty} h(t)\, dt = \tau \tag{2.2.15}$$

$$|H(f)|^2\,df = \int_{-\infty}^{\infty} h^2(t)\,dt = \tau/2 \tag{2.2.16}$$

In this case,

$$\Delta T_{rms} = \frac{\left(T_A + T_R\right)}{\sqrt{2B\tau}} \tag{2.2.17}$$

Although this filter provides slightly greater sensitivity, it is at the expense of a longer memory, so large transients may take several time constants to fade away.

Such total power radiometers can be quite sensitive. Consider a typical radio astronomy receiver looking at cold sky where $T_A + T_R = 30K$; for a 100-MHz bandwidth B and 1-second integration the sensitivity is:

$$\Delta T_{rms} = 30 \Big/ \sqrt{10^8 \bullet 1\ \text{sec}} = 0.003\ \text{K} \tag{2.2.18}$$

This sensitivity can be improved by a factor of 10 if we average the data for 100 seconds.

A contrasting example for which very poor sensitivity suffices is an amplitude modulated (AM) radio for which low-cost electronics readily provide effective system temperatures of 10,000K or better, over nominal bandwidths B of 10 kHz for time constants τ of ~10^{-4} seconds so that:

$$\Delta T_{rms} = 10^4 \Big/ \sqrt{10^4 \bullet 10^{-4}} = 10^4 \text{K} \tag{2.2.19}$$

Even though this sensitivity appears quite poor, it is often overwhelmed by interference from adjacent AM broadcasters, or static from lightning or household appliances. Thermal noise is the sound we hear on AM radios when we are tuned between stations and free from station interference.

2.2.2 Measurement of thermal power using a sampled system

A more intuitive understanding of the sensitivity equation (2.2.14) can be derived by analyzing the sampled-pulse version of the same total-power radiometer, as suggested in Figure 2.2-7.

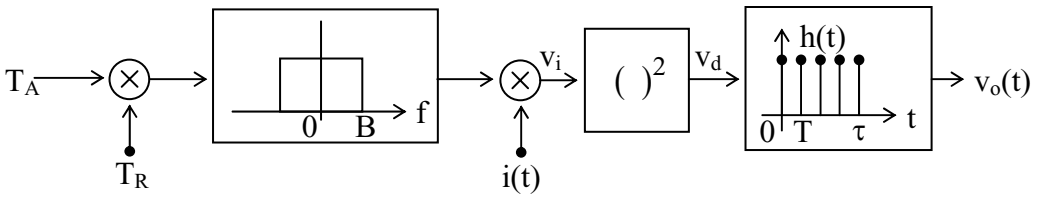

Figure 2.2-7 Architecture of a sampled total-power radiometer.

The sampling impulses i(t) are spaced at intervals of T seconds, where T = 1/2B is the Nyquist sampling rate. If T < 1/2B, then the sampling theorem shows that adjacent pulses will have correlated amplitudes, and if T > 1/2B, some of the information in the incoming signal T_A will be lost. Note that if the signal extends from zero Hertz then Nyquist sampling corresponds to two samples per period of the highest frequency component present. The boxcar averages $\tau/T = 2B\tau$ pulses, each of amplitude v_d. If this number $2B\tau$ is much larger than 1, then the output voltage $v_o(t)$ approaches a Gaussian distribution by the central limit theorem, and the variance of the output voltage is:

$$\text{Variance of } v_o = 2B\tau\sigma_d^2 \tag{2.2.20}$$

The variance σ_d^2 of the detected voltage v_d is:

$$\sigma_d^2 \triangleq \overline{\left(v_d - \overline{v_d}\right)^2} = \overline{\left(v_i^2 - \overline{v_i^2}\right)^2} = \sigma_d^2 = \overline{v_i^4} - 2\left(\overline{v_i^2}\right)^2 + \left(\overline{v_i^2}\right)^2 = \overline{v_i^4} - \left(\overline{v_i^2}\right)^2 \tag{2.2.21}$$

We may recall that if x is a jointly Gaussian random variable of zero mean, then:

$$\overline{x^n} = 1 \bullet 3 \bullet 5 \bullet\bullet (n-1), \text{ if n even; } \overline{x^n} = 0, \text{ if n odd} \tag{2.2.22}$$

If we define the constant a such that:

$$\overline{v_i^2} = aT_{eff}\,\overline{x^2} \text{ and } \overline{x^2} \equiv 1 \tag{2.2.23}$$

then (2.2.21) becomes:

$$\sigma_d^2 = \overline{v_i^4} - \left(\overline{v_i^2}\right)^2 = T_{eff}^2 a^2 \left[\underbrace{\overline{x^4}}_{3} - \left(\underbrace{\overline{x^2}}_{1}\right)\right] = 2T_{eff}^2 a^2 \tag{2.2.24}$$

The variance of the output voltage v_o follows from (2.2.20) and (2.2.24):

$$\text{variance of } v_o = 2B\tau 2T_{eff}^2 a^2 \tag{2.2.25}$$

It can be seen from Figure 2.2-7 that the average value of $\bar{v}_o$ of the output voltage is:

$$\bar{v}_o = 2B\tau \bullet \overline{v_i^2} = 2B\tau \bullet T_{eff} a \tag{2.2.26}$$

Therefore it follows from (2.2.21), (2.2.25), and (2.2.26) that:

$$\Delta T_{rms} = \frac{\sqrt{\text{variance of } v_o}}{\partial \bar{v}_o / \partial T_A} = \frac{T_{eff} a \sqrt{4B\tau}}{2B\tau a} = T_{eff} / \sqrt{B\tau} \tag{2.2.27}$$

This result, which is the same as (2.2.14), now has a more intuitive interpretation, where the rms sensitivity approximately equals the fluctuations associated with T_{eff} divided by the square root of the number of independent samples that were averaged, $2B\tau$.

2.2.3 Power measurement errors due to gain fluctuations, and the remedies

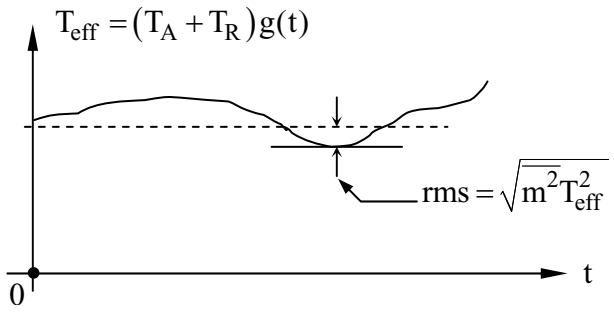

Figure 2.2-8 Effects of gain fluctuations on the output of total-power radiometers.

Many total power radiometers have sensitivities far worse than those suggested by (2.2.27) as a result of gain fluctuations g(t) in the amplifier train. This problem is illustrated in Figure 2.2-8, where the receiver output $v_o(t)$, which is proportional to T_{eff}, varies due to gain fluctuations g(t) to a degree that overwhelms the thermal noise.

A simple expression for receiver sensitivity in the presence of gain fluctuations results when the receiver gain can be approximated as:

$$g(t) \cong G(1 + m(t)), \ |m| << 1 \tag{2.2.28}$$

Since the gain fluctuations are generally independent of the thermal noise, the variances add so that the effective receiver sensitivity is:

$$\Delta T_s \cong \sqrt{\left(\Delta T_{thermal}\right)^2 + \overline{m^2} T_{eff}^2} = T_{eff} \sqrt{\frac{1}{B\tau} + \overline{m^2}} \tag{2.2.29}$$

The potential seriousness of gain fluctuations is suggested by noting that a 0.1-percent gain fluctuation for a system having T_{eff} of 1000K is 1K, large compared to the 3-mK sensitivity suggested for the example given in Equation (2.2.18). Gain fluctuations are particularly prevalent in the very high gain amplifier chains employed in high sensitivity receivers, where these gains can exceed 100 dB. Gain fluctuations often arise from small thermal fluctuations or from semiconductor instabilities.

A simple widely used remedy for gain fluctuations is called "synchronous detection," and is generally implemented as suggested in Figure 2.2-9.

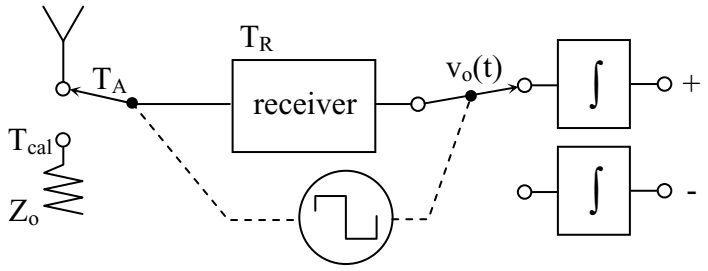

Figure 2.2-9 Synchronous detection in a total-power radiometer.

Such synchronous detectors are often called Dicke radiometers, named after their developer at the MIT World War II Radiation Laboratory. The upper integrator is connected to the output voltage $v_o(t)$ every time the input switch is connected to the antenna. A square-wave generator moves the pair of switches synchronously between their upper and lower positions at both input and output. If the calibration temperature T_{cal} is slightly different from the antenna temperature T_A, then the output voltage exhibits the form shown in Figure 2.2-10.

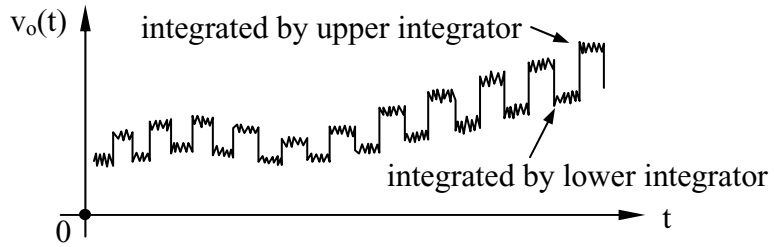

Figure 2.2-10 Receiver voltage prior to the output synchronous detector switch.

In the special case where the input is balanced so that $T_{cal} = T_A$, then the two output integrators will have the same average voltage and the average (noise-free) output will be zero, independent of gain fluctuations in the receiver. Any such gain fluctuations can operate only on the difference between T_A and T_{cal}, and not on the total T_{eff}. The rms fluctuations of this difference between the two output integrators is independent of the way in which the output switch is operated, provided the switch is operated sufficiently rapidly that the signal $v_o(t)$ is highly correlated between time intervals. Although the thermal contribution to the output fluctuations is unchanged by splitting the output integrator and adding a switch, a 50-percent duty cycle results in the antenna being observed only half the time. As a result the output signal amplitude is reduced by a factor of two, and the denominator of (2.2.1) is also reduced by a factor of two, resulting in:

$$\Delta T_{rms_{Dicke}} = 2T_{eff} \Big/ \sqrt{B\tau} \tag{2.2.30}$$

This factor-of-two penalty associated with a 50-percent duty cycle synchronous detector can be reduced by observing the reference signal less than half the time. To provide equivalent smoothing of the reference signal at the output, however, the integration time of only that integrator should be lengthened appropriately. The gains associated with the two integrators may have to be adjusted also, depending on the choices of these two integrations times. Both integration times should be shorter than the characteristic fluctuation time constant for the amplifier gain and the desired signal, but much longer than the period of the synchronous detector.

For receivers with very large bandwidths the variance of the output signal $v_o(t)$ shown in Figure 2.2-9 is sometimes sufficiently large that it is useful to reduce it to prevent saturation of the output amplifier. One way to do this is to insert a narrowband filter tuned to the Dicke switch frequency, typically 1 kHz, but this also exacts a small penalty because the desired square-wave signal is also partially filtered out, resulting in:

$$\Delta T_{rms} = \frac{\pi}{\sqrt{2}} T_{eff} \Big/ \sqrt{B\tau} \tag{2.2.31}$$

2.2.4 Correlation receivers

Another widely used receiver configuration is that of the correlation radiometer, illustrated in Figure 2.2-11, which is used in interferometers, total-power radiometers, and in matched-filter receivers used for communications and radar. As shown in the figure, the correlation radiometer can avoid the effects of gain fluctuations because the average product of the two independent amplifier noises n_a and n_b is zero, so gain fluctuations cannot impact their contribution. If the signals arrived from two independent antennas, then such a system can be used as an interferometer as discussed later. If the incoming signal from the antenna is fed only to the first amplifier, and the signal from the other side is noise free, then the resulting analysis is that of the matched-filter receiver, which is also discussed later.

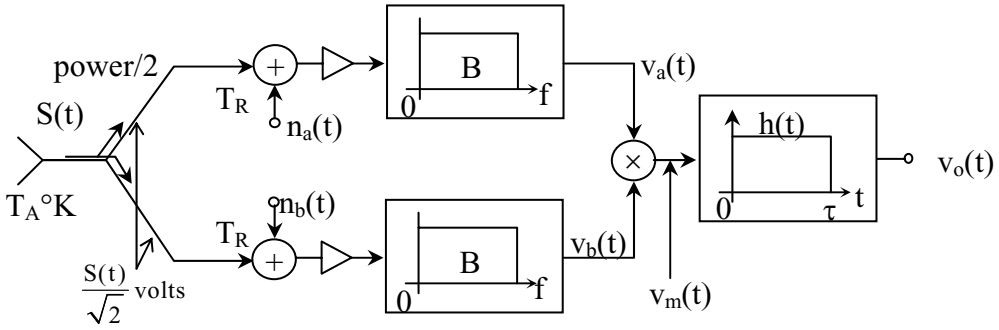

Figure 2.2-11 Architecture of a correlation radiometer.

The analysis below follows that of the total power radiometer with only minor variations. We may begin by computing the autocorrelation function of the multiplier output $\phi_m(t)$:

$$\begin{aligned}\phi_m(\tau) &= E\left[v_a(t)\,v_b(t)\,v_a(t-\tau)\,v_b(t-\tau)\right] \\ &= E\left[\left(\frac{S_1}{\sqrt{2}}+n_{a_1}\right)\left(\frac{S_1}{\sqrt{2}}+n_{b_1}\right)\left(\frac{S_2}{\sqrt{2}}+n_{a_2}\right)\left(\frac{S_2}{\sqrt{2}}+n_{b_2}\right)\right]\end{aligned} \tag{2.2.32}$$

Equation (2.2.32) employs a simplified notation where the subscripts 1,2 refer to the times t and t - τ, respectively. Note that the incoming signal power associated with S(t) is divided in two by a matched power divider. This implies that the voltage in each path is reduced by $1/\sqrt{2}$, assuming the impedances of the two outputs are the same as that of the input line. Because the four variables multiplied in (2.2.32) are jointly Gaussian random variables of zero mean, we may again use (2.2.3) to yield:

$$\phi_m(\tau) = \frac{1}{4}\phi_s^2(0) + \frac{1}{2}\phi_s^2(\tau) + \phi_s(\tau)\phi_n(\tau) + \phi_n^2(\tau) \tag{2.2.33}$$

The power density spectrum $\Phi_m(f)$ can be found from the Fourier transform of (2.2.33):

$$\Phi_m(f) = \frac{1}{4}\phi_s^2(0)\delta(f) + \frac{1}{2}\Phi_s(f) * \Phi_s(f) + \Phi_s(f) * \Phi_n(f) + \Phi_n(f) * \Phi_n(f) \qquad (2.2.34)$$

The DC power is associated with $\phi_s^2(0)\delta(f)$ and the AC power is associated with the other terms in (2.2.34). Using the methods in (2.2.11) and (2.2.12), we can use (2.2.1) to show

$$\Delta T_{rms} = \frac{\sqrt{P_{ac}}}{\partial\sqrt{P_{dc}}/\partial T_A} = \frac{T_{eff}}{\sqrt{B\tau}} \qquad (2.2.35)$$

where:

$$T_{eff}^2 = T_A^2 + 2T_A T_R + 2T_R^2 \qquad (2.2.36)$$

The rms sensitivity of Equation (2.2.35) reduces to $\sqrt{2}T_R/\sqrt{B\tau}$ for the weak-signal case where $T_A << T_R$, and to the limit $T_R/\sqrt{B\tau}$ for the strong-signal case where $T_A >> T_R$.

2.2.5 Measurement of power spectra

Such total power radiometers can also be combined or reconfigured to measure power spectra of unknown signals. In general, spectral analysis is obtained by splitting the bandwidth into multiple frequency bands shaped by individual filters, and then detecting each band separately. The receiver configuration employed typically depends on the total bandwidth, the number of spectral channels desired, and the absolute frequencies to be monitored.

The most extreme case is that where the bandwidth to be observed exceeds that of available antennas or amplifiers. In this case separate systems are required. More often the antenna bandwidth is adequate but that of the amplifiers is not. In this case passive frequency dividers are used between the antenna and the amplification or detection stage. If amplifiers of adequate bandwidth are available, the signals are generally amplified before they are detected or split further. In rare cases the signal is sufficiently large compared to detector noise that amplifiers can be omitted. If the bandwidth is sufficiently narrow, digital spectral analysis can be employed, as discussed later.

The most important property of passive filters used for spectral measurements is that they be low-loss to minimize their contributions to Johnson noise, and that they have appropriately shaped frequency bandpass characteristics.

Figure 2.2-12 shows a typical channel-dropping RLC filter chain. The operation of the RLC filter circuit in Figure 2.2-12 is easily understood. If each of the series and parallel resonators is

nearly lossless (high-Q), then at frequencies sufficiently remote from any resonance f_i the parallel resonators become short-circuits and the series resonators become open circuits, thereby connecting the input directly to the output load Z_o.

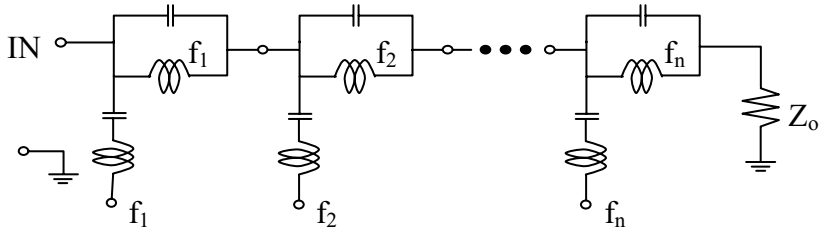

Figure 2.2-12 Multistage passive channel-dropping RLC filter chain.

At the resonant frequency f_2 the associated parallel resonator becomes an open circuit, thereby shunting the input signal through the series resonator to the output port for f_2, where the series resonator approximates a short circuit. The same is true for any of the other frequency taps in the chain, and the separate channel dropping filters can be connected in any order. They begin to interfere principally when the frequency bands begin to overlap significantly.

The same concept can also be applied to chains of waveguide filters as suggested in Figure 2.2-13.

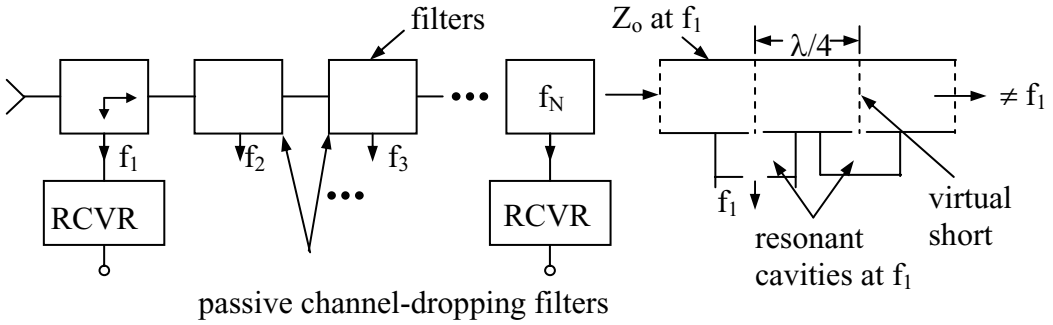

Figure 2.2-13 Channel-dropping filters implemented using waveguides.

When the second cavity is resonant at f_1 the field amplitudes inside build to levels sufficient to produce a virtual short circuit in the plane of the aperture connecting the wave guide to the cavity: one-quarter wavelength down the waveguide this appears as an open circuit, analogous to the behavior of the parallel resonator in the RLC channel-dropping filter system. If the first cavity is resonant at this same frequency f_1 and if both its input and output apertures exhibit the same external Q, then all of the power input at f_1 can be shunted to a matched load at the filter output.

Standard frequency division techniques used at infrared and visible frequencies are suggested in Figure 2.2-14, including prisms, diffraction gratings, and cascaded dichroic beamsplitters.

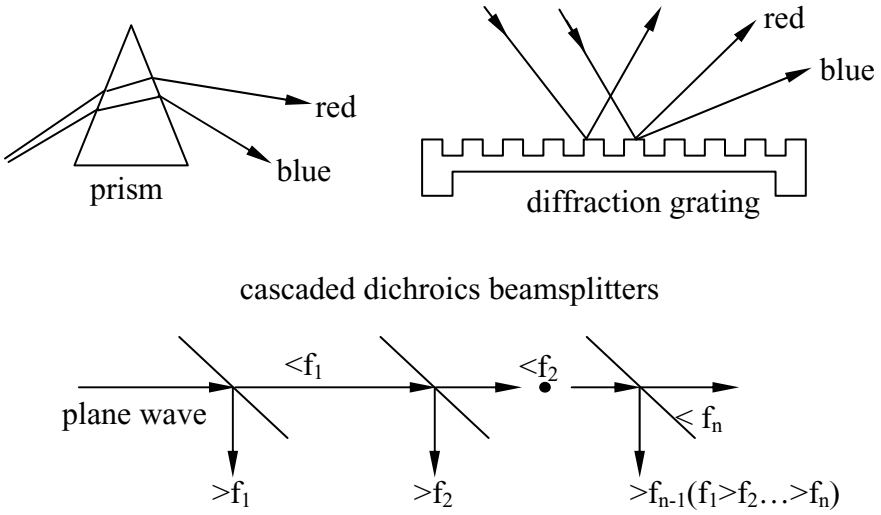

Figure 2.2-14 Channel-dropping filters implemented using a prism, diffraction grating, and cascaded dichroic beamsplitters.

Most materials like glass or plastic are dispersive as a result of the frequency-dependent permittivity associated with bound electrons. For example, blue light typically interacts more with these electrons and is refracted more than red light. Diffraction gratings having lines ruled with separations on the order of a wavelength can diffract different frequencies in different directions with high efficiency and low loss, and are commonly used at infrared wavelengths where prisms are not available. Diffraction gratings become highly dispersive if the lines are ruled at extremely close spacings. Dichroic mirrors typically pass frequencies above or below some cutoff frequency and reflect the rest of the spectrum. In Figure 2.2-14 all frequencies above f_1 are reflected to one side by the first dichroic mirror, while those above f_2 but below f_1 are shunted aside by the second mirror. Because the insertion loss of such dichroic mirrors can be large, typically no more than a few are cascaded at any one time. For larger numbers of channels diffraction gratings are usually used.

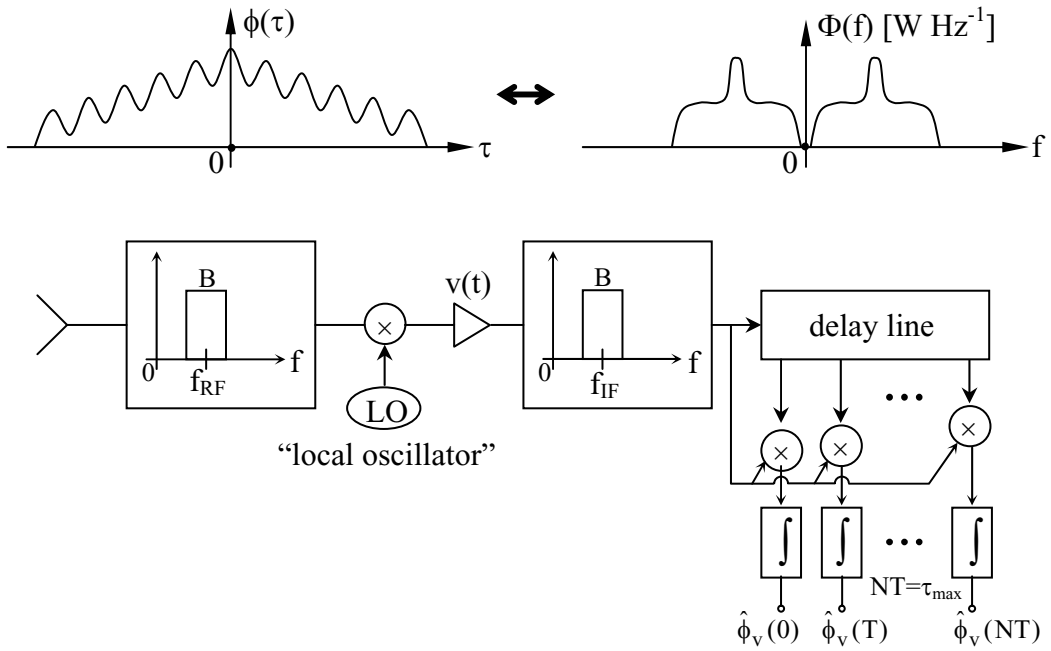

Figure 2.2-15 System for spectral analysis using analog autocorrelation and delay lines.

Digital spectral analysis is possible using conventional Fourier transform chips manufactured at low cost for digital signal processing (DSP) purposes. We can also use chips that compute related transforms, such as the discrete-cosine transform, and others. In general the bandwidths of the frequency channels extracted in this fashion are a small fraction of the clockspeed of the chip because of the large number of computations required. Highly parallel structures can increase both the number of channels and their bandwidths, but the clockspeed of the circuit remains a significant barrier.

For many purposes much less expensive digital circuits can be employed. To understand them, first consider the use of autocorrelation functions as an intermediate computational step in an analog spectral analysis system, as suggested in Figure 2.2-15. The illustrated circuit first defines an overall bandwidth B which is to be further divided into spectral channels by autocorrelation analysis. The local oscillator and preamplifier translate this band to low frequencies where it can be introduced to a delay line taped uniformly along its length at N points, where the maximum delay is $NT = \tau_M$. The resulting spectral resolution and coverage are limited by the delay line length and number of taps. The system in Figure 2.2-15 then computes the autocorrelation function $\phi(\tau)$ for an integration time much longer than τ_M. A separate computer transforms this output into $\Phi(f)$, typically in units of watts/Hertz.

Consider first the effects of the finite length of the delay line, where the maximum delay $\tau_M = NT$. The observed autocorrelation function $\hat{\phi}_v(\tau)$ is then the true autocorrelation

function $\phi_v(\tau)$ times the weighting function $W(\tau)$, which consists of a boxcar of value unity for delays τ ranging between $-\tau_M$ and $+\tau_M$. This boxcar weighting function controls the spectral response of this autocorrelation system as suggested in equation (2.2.36).

$$\begin{array}{ccccc} \hat{\phi}_v(\tau) & = & \phi_v(\tau) & \bullet & W(\tau) \\ \updownarrow & |\tau| < \tau_M & \updownarrow & \updownarrow & \updownarrow \\ \hat{\Phi}_v(f) & = & \Phi_v(f) & * & W(f) \end{array} \qquad (2.2.36)$$

Figure 2.2-16 illustrates this spectral response, where the first null appears at $1/2\tau_M$ Hz from the center frequency of each channel and the response is a sinc function with rather high sidelobes.

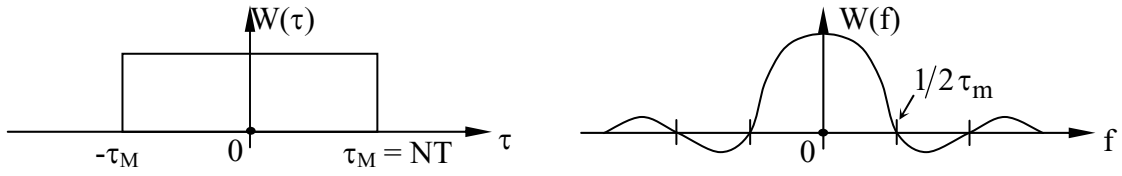

Figure 2.2-16 Spectral response of an autocorrelation receiver with a finite-length delay line.

The consequences of sampling an infinite delay line by an impulse train i(t) with a period of T seconds are suggested by equation (2.2.37).

$$\begin{array}{ccccc} \hat{\phi}_v(\tau) & = & \phi_v(\tau) & \bullet & i(t) \\ \updownarrow & & \updownarrow & \updownarrow & \updownarrow \\ \hat{\Phi}_v(f) & = & \Phi_v(f) & * & I(f) \end{array} \qquad (2.2.37)$$

Because the true autocorrelation function $\phi_v(f)$ is sampled by the impulse train i(t), the resulting estimated power density spectrum $\hat{\Phi}_v(f)$ is the convolution of the true power spectral density $\hat{\Phi}_v(f)$ with the transformed impulse train I(f), a train of spectral impulses with spacing 1/T Hz. Unless the delay line is sampled sufficiently frequently, the estimated power density spectrum can be aliased as suggested in Figure 2.2-17.

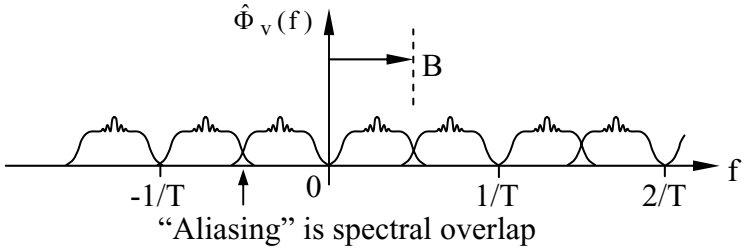

Figure 2.2-17 Aliasing in autocorrelation receivers for $1/T \cong 2B$.

Thus, to avoid aliasing, the larger the bandwidth B to be analyzed, the shorter the time delay T between successive taps on the delay line must be.

To reduce the sidelobes in the spectral response associated with the weighting function $W(\tau)$ and finite delay line length, it is common to apodize the weighting function by rounding off its corners. For example, one common apodization is to replace the boxcar by the positive half of a cosine function. Because the autocorrelation function is averaged for only a limited time before the Fourier transform $\Phi_v(f)$ is computed, both the autocorrelation function and the computed spectrum have additive noise, but this is generally equivalent to the noise incurred by comparable RLC filters acting on finite-duration random signals.

2.2.6 Signal processing with clipped signals: autocorrelators

As a practical matter, arrays of analog correlators are never used because it is difficult to implement high quality analog multipliers without offsets and nonlinearities, which become important as integration times are increased. Such correlators have found wide application, however, in digital systems where the signals are reduced to 1-bit hard clipped approximations before entering the delay line, as suggested in Figure 2.2-18. Although more bits can be used, 1-bit or 2-bit systems offer enormous savings in hardware complexity with little loss in noise performance.

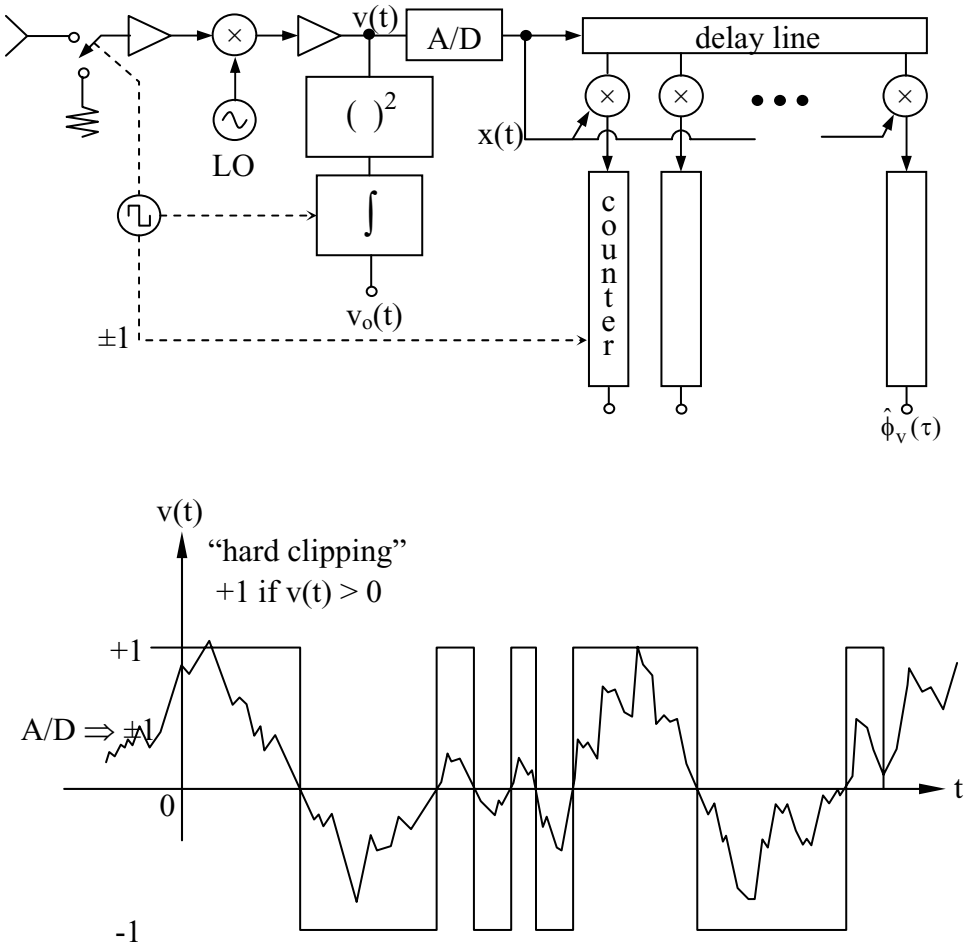

Figure 2.2-18 Autocorrelation spectral analysis system for hard clipped signals.

Figure 2.2-18 suggests how hard clipping reduces the signal to a string of ones and zeros, or to +1 and -1. We show below that the autocorrelation function for this clipped signal is related in a simple way to the true spectrum. It also can be shown that the penalty in rms system sensitivity is no more than ~50 percent ($\sigma_{clip} \cong 1.5\ \sigma_{optimum}$), and use of two bits instead of one reduces this small penalty further. Because such clipping removes information about the total received power, it is necessary to have a separate total power measurement $v_o(t)$, as suggested in the figure. Because the delay line is now handling one-bit signals, it can be implemented as a simple shift register, and each multiplier can be implemented with only a few logic gates. The integrators become simple counters. As a result very powerful digital correlation receivers can be implemented very compactly on silicon. Autocorrelation functions for signals with multi-gigahertz bandwidths can be computed in real time today.

That autocorrelation functions comprising averaged products of one-bit clipped signals are simply related to the power spectra of the original unclipped signals can be readily demonstrated.

The following suggests the general approach. Let the input voltage at time t_1 be $x(t_1) = x_1$, and sgn x be defined as +1 for $x \geq 0$, and -1 when $x \leq 0$, where x_1 and x_2 are jointly Gaussian random variables of zero mean. Then the autocorrelation function of the signal x(t) is:

$$\phi_x(\tau) = E[\text{sgn } x_1 \text{ sgn } x_2] = \int_{-\infty}^{\infty}\int \text{sgn } x_1 \text{ sgn } x_2 \left[\frac{1}{2\pi(1-\rho)^{1/2}} e^{-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}}\right] dx_1 dx_2 \qquad (2.2.38)$$

where $\tau = t_2 - t_1$ and $\rho = \overline{x_1 x_2} = \phi_v(\tau)$. Considering the four possible combinations for the product sgn x_1 sgn x_2, we obtain:

$$\begin{aligned}\phi_x(\tau) &= 2\int_0^{\infty}\int [p(x_1, x_2)]dx_1 dx_2 - 2\int_{-\infty}^{0}\int_0^{\infty} p(x_1, x_2)dx_1 dx_2 \\ &= 4\int_0^{\infty}\int p(x_1, x_2)dx_1 dx_2 - 1\end{aligned} \qquad (2.2.39)$$

where we note that the second integral term in (2.2.39) equals twice the first integral minus one. By converting (2.2.38) to circular coordinates the exponent can be simplified, permitting (2.2.39) to be evaluated:

$$\hat{\phi}(\tau) \equiv \hat{\rho} = \sin\left(\frac{\pi}{2}\hat{\phi}_x(\tau)\right) \qquad (2.2.40)$$

where the estimated autocorrelation function $\hat{\phi}_x(\tau)$ is computed for T seconds.

The estimate $\hat{\phi}(\tau)$ is increasingly biased as $\hat{\phi}_x(\tau)$ becomes less accurate because $\hat{\phi}(\tau)$ is a nonlinear function of the observed $\hat{\phi}_x(\tau)$. For the special case where there are N delay-line taps and we use uniform boxcar weighting of the autocorrelation function together with spectral samples spaced at intervals of $1/2\tau_M$ Hz, we can show that the number of independent spectral samples equals N. In practice the number of delay line taps might be twice the number of spectral samples to account for apodization of the weighting function $W(\tau)$ and to avoid aliasing associated with the skirts of the lowpass filter that defines the bandwidth B.

2.3 POWER AND NOISE PROPAGATION IN RECEIVERS

2.3.1 Power and gain in circuits

The building blocks comprising receivers are generally characterized in terms of their gain and noise figure, whereas signals are generally characterized in terms of their power spectral density and signal-to-noise ratio. Relations governing these quantities in receivers are discussed here, together with circuits for canceling unwanted noise and interference.

To begin, it is important to distinguish between different possible definitions for power observed at the junction between a circuit and its load $\underline{Z}_L$. Figure 2.3-1 defines such a junction driven by a *Thevenin equivalent circuit*, for the generator which is characterized by its open circuit voltage V_g and its *source impedance* $\underline{Z}_g = R_g + jX_g$.

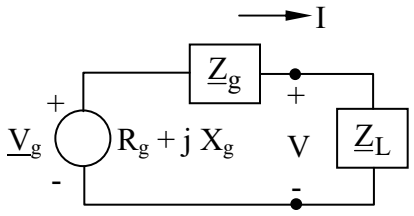

Figure 2.3-1 Thevenin equivalent circuit for a generator driving a load.

If we characterize the voltage v(t) at this junction for a monochromatic signal at frequency ω, where $v(t) = R_e\{\underline{V}e^{j\omega t}\}$, and a similar definition is used for a phasor $\underline{I}$ representing the current flowing into the load, then we may simply characterize three types of power:

$$P_{delivered} \triangleq \frac{1}{2} R_e\{\underline{V}\underline{I}^*\} \qquad (\triangleq P_D) \tag{2.3.1}$$

$$P_{available} \triangleq \max\ P_D, \text{ i.e. if } \underline{Z}_L = \underline{Z}_g^* \qquad (\triangleq P_A) \tag{2.3.2}$$

$$P_{exchangeable} \triangleq |P_D|_{\underline{Z}_L = \underline{Z}_g^*} \qquad (\triangleq P_E) \tag{2.3.3}$$

The *delivered power* P_D is the power actually delivered to the load by the particular source that is present. In contrast, the *available power* P_A depends only on the source because we assume that the load is matched; P_A is the maximum power extractable from the given Thevenin equivalent source. Figure 2.3-2 suggests the dependence of this delivered power P_D as a function of load resistance R_L.

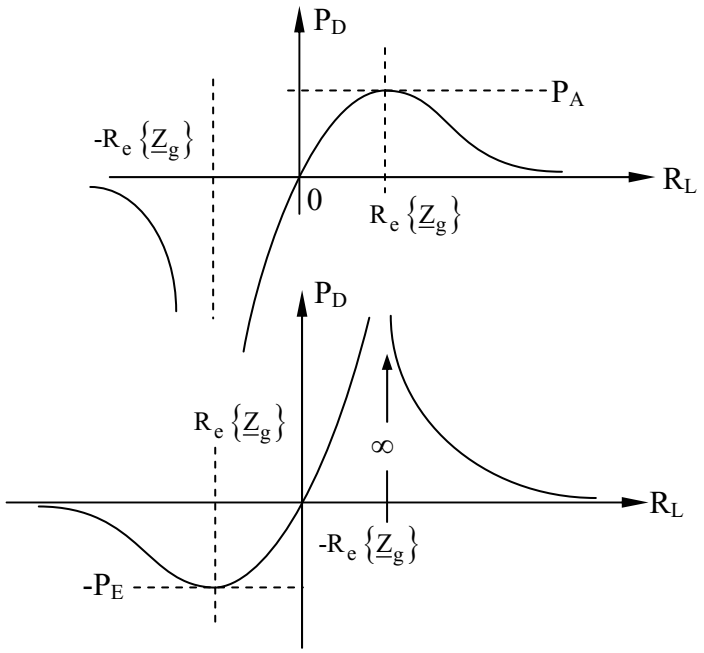

Figure 2.3-2 Dependence of delivered power P_D on load resistance R_L for a positive and a negative source impedance Z_g.

Figure 2.3-2 shows that P_D reaches a maximum value P_A when $R_L = R_e\{\underline{Z}_g\}$. If a negative load resistor R_L is connected to the source such that $R_L = -R_e\{\underline{Z}_g\}$, then the power flow from the load to the source becomes infinite. Conversely, if the source has a negative impedance, then it can deliver infinite power to a matched load of positive impedance. This issue is not purely academic, because many amplifiers are implemented by reflecting incoming signals from elements with negative resistances near their operating points; examples include *parametric amplifiers* and some masers. Because *negative-resistance amplifiers* are generally used near their stable operating point, we also define *exchangeable power* P_E as the magnitude of P_D for the case where the load is a conjugate match to the generator, as suggested in (2.3.3).

Consider now an amplifier of gain G which connects a generator of impedance $\underline{Z}_g$ to a load $\underline{Z}_L$. If we represent the input and output terminals of the amplifier by the subscripts 1 and 2, respectively, then we can simply define several different types of gain:

$$G_{\text{insertion}}\,(=G_I) \triangleq \frac{P_{D_2}\ (\text{with amplifier in})}{P_{D_2}\ (\text{without amplifier})} \tag{2.3.4}$$

$$G_{\text{power}}\left(=G_p\right) \triangleq \frac{P_{D_2}}{P_{D_1}} \tag{2.3.5}$$

$$G_{\text{available}}\left(=G_A\right) \triangleq \frac{P_{A_2}}{P_{A_1}} \tag{2.3.6}$$

$$G_{\substack{\text{transducer}\\ \text{power}}}\left(=G_T\right) \triangleq \frac{P_{D_2}}{P_{A_1}} \tag{2.3.7}$$

$$G_{\text{exchangeable}}\left(=G_E\right) \triangleq \frac{P_{E_2}}{P_{E_1}} \tag{2.3.8}$$

These definitions all yield the same gain if source, amplifier, and load are all matched to one another and have positive real-part impedances. The distinction becomes important only when systems are mismatched, in which case the preferred definition of gain is G_E, or its positive-resistance equivalent, P_A. The advantage of these definitions for G_A and G_E is that they explicitly do not depend on the load impedance $\underline{Z}_L$, although they do depend on the generator $\underline{Z}_g$. Hereafter we shall use only exchangeable power and gain in our discussion.

2.3.2 Noise figure

Consider the amplifier illustrated in Figure 2.3-3. We can characterize the amplifier in Figure 2.3-3 in terms of its gain G and its noise figure F, where the noise figure F can be defined in terms of the signal-to-noise ratios at the input terminals (1) and the output terminals (2). These signal-to-noise ratios are:

$$\text{SNR}_1 \triangleq S_1/N_1, \text{ and } \text{SNR}_2 \triangleq S_2/N_2 \tag{2.3.9}$$

where we define N_1 as the exchangeable noise power spectrum at the input (1) and N_2 as the same, but at the output (2). Similarly, we define S_1 as the exchangeable signal power spectrum at the input port (1), and S_2 as the same, but at the output port (2).

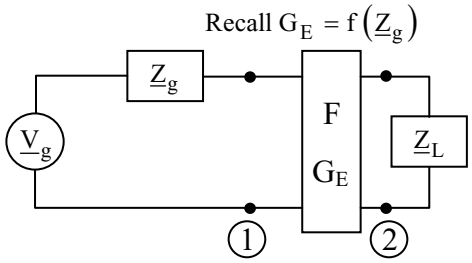

Figure 2.3-3 Amplifier embedded in a linear circuit.

We can now define the *noise figure* F:

$$F \triangleq \frac{SNR_1}{SNR_2} \equiv \frac{S_1/N_1}{S_2/N_2},\ N_1 \triangleq kT_o,\ T_o \triangleq 290K \tag{2.3.10}$$

This definition governs commercial products and has legal force; see page 436 of the *Proceedings of the IEEE* for March, 1963. Note that this definition explicitly employs an input noise temperature N_1 of 290 K, regardless of the circumstances under which the amplifier might be used in practice.

In general, the output noise from an amplifier N_2 is larger than the amplified input noise GN_1 by an amount N_{2T} introduced by the transducer, which we call *transducer noise*. The noise figure F can be simply related to N_{2T} by using the definition (2.3.10):

$$F = \frac{S_1/N_1}{GS_1/(GN_1 + N_T)} = 1 + \frac{N_{2T}}{N_1 G} \tag{2.3.11}$$

Sometimes we characterize an amplifier by its *excess noise figure*, defined as F-1, where:

$$F - 1 = \frac{N_{2T}}{N_1 G} \triangleq \frac{kT_R G}{kT_o G} = \frac{T_R}{T_o} \tag{2.3.12}$$

where (2.3.12) also defines the *receiver noise temperature* T_R, which can be interpreted as suggested in Figure 2.3-4.

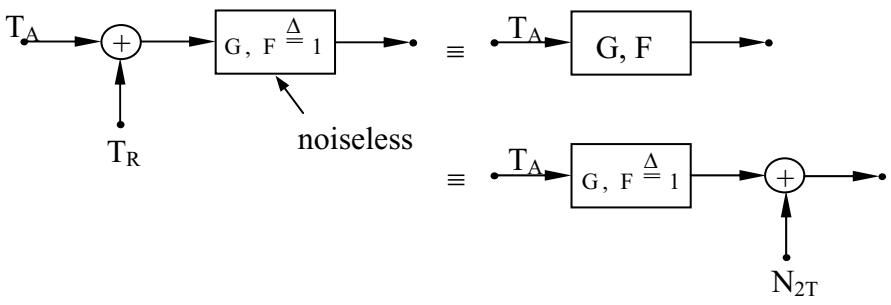

Figure 2.3-4 Noisy amplifier: three equivalent circuits.

The relationships defined by (2.3.12) are illustrated in Figure 2.3-4. A noisy amplifier with input T_A, gain G, and noise figure F can be represented in three different ways, as illustrated in Figure 2.3-4. First, a noisy amplifier can be represented as a noiseless one preceded by addition of the equivalent receiver noise temperature T_R (°K). Noise temperatures of cryogenically cooled amplifiers are typically ~5-50K up to 10 GHz or more, and ~100-1000K for uncooled systems. Alternatively, the same noisy amplifier can be represented by a noiseless system followed by addition of the *transducer noise* N_{2T}. Equation 2.3.12 quickly relates receiver noise temperatures T_R to the corresponding noise figure F; for example, receiver noise temperatures of 0K, 290K, and 1500K correspond to noise figures F of 1, 2, and approximately 6, respectively. Noise figures are often expressed in units of dB, so these three noise figures can also be characterized as 0, 3, and approximately 7.5 dB.

It is often useful to know the effective noise figure and gain of two or more amplifiers connected in series. Figure 2.3-5 shows how an amplifier system of noise figure F_{1+2} and gain G_{1+2} results when an amplifier characterized by F_2, G_2 follows an amplifier characterized by F_1, G_1. The noise figure F_{1+2} follows from its definition:

$$F_{1+2} \triangleq \frac{S_1/N_1}{S_3/N_3} \tag{2.3.13}$$

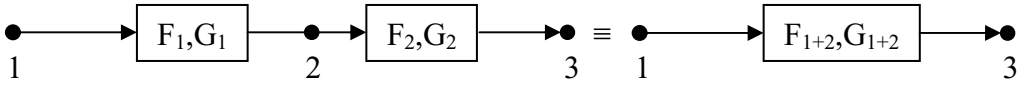

Figure 2.3-5 Characterization of two cascaded amplifiers.

Our previous definitions for gain imply that the output signal power spectrum $S_3 = G_1 G_2 S_1$. The noise power spectrum N_2 at the output of the first amplifier equals the noise power spectrum at

the input of the amplifier, kT_oF_1, times the gain G_1 of the first amplifier; this reflects the fact that noise figure definitions assume kT_o is input to the amplifier, where T_o is defined as 290K.

The noise power spectrum N_3 at the output of the cascade has contributions from N_2 amplified by G_2 and from the excess noise introduced by amplifier 2:

$$N_3 = G_2\ kT_oF_1G_1 + G_2\left(F_2 - 1\right)kT_o \tag{2.3.14}$$

It follows from (2.3.14) that:

$$S_3/N_3 = G_1G_2S_1/\left[\left(G_1G_2F_1 + G_2\left(F_2 - 1\right)\right)kT_o\right] = \left(S_1/kT_o\right)/\left(F_1 + \left(F_2 - 1\right)/G_1\right) \tag{2.3.15}$$

Equation (2.3.13) can be combined with (2.3.15) and the fact that $kT_o = N_1$ to yield:

$$F_{1+2} = \frac{S_1/N_1}{S_3/N_3} = F_1 + \frac{F_2 - 1}{G_1} \tag{2.3.16}$$

By extension we may readily find expressions for longer cascades of amplifiers. For example, a cascade of three amplifiers can be handled by first cascading the first two, yielding an amplifier characterized by F_{1+2}, as follows:

$$F_{1,2,3} = F_{1+2} + \frac{F_3 - 1}{G_1G_2} \tag{2.3.17}$$

We may iteratively combine (2.3.16) and (2.3.17) to find a general expression for cascaded amplifiers having any number of stages:

$$F_{1,2,\ldots} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1G_2} + \ldots \tag{2.3.18}$$

It is important to note that it is not obvious which sequence of two amplifiers will yield better noise performance. Equation 2.3.16 shows that the combined noise figure F_{1+2} depends on both the noise figure F_1 and the gain G_1 of the first amplifier, as well as the noise figure F_2 of the second. Only by evaluating both F_{1+2} and F_{2+1} can we be certain which sequence is preferable. In general the lowest combined noise figure results when the amplifier with the lower noise figure is placed first, unless that amplifier has much lower gain than the other.

2.3.3 Superheterodyne circuits

An important element in many receivers is a *superheterodyne* stage in which the signal from a *local oscillator* "L.O." multiplies a signal before it is amplified at the *intermediate frequency* "i.f. frequency", which is the difference between the signal and L.O. frequencies. A combination of the L.O. and the multiplier is often called a "*mixer*", characterized by G_c, F_c. The system consists of an input bandpass filter which selects the frequency band to be down-converted; this filter precedes the mixer and the following *i.f. preamplifier*, which amplifies the signals within the much lower-frequency i.f. passband, which is defined by the i.f. passband filter, all as illustrated in Figure 2.3-6.

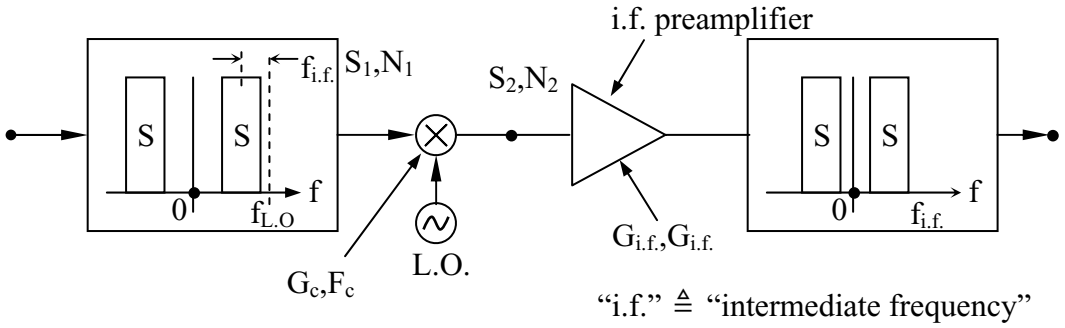

Figure 2.3-6 Superheterodyne circuit.

The frequency relationships between the various passbands in a superheterodyne system are suggested in Figure 2.3-7. They result from the fact that multiplication of signals in the time domain is equivalent to convolution in the frequency domain. Therefore the signal spectrum S_1 input to the mixer is translated by the superheterodyne to the i.f. passband where it becomes the signal spectrum S_2. Alternatively, the input (often at radio frequencies, or "r.f.") can be at frequencies above the local oscillator frequency f_{LO}. Since the mixer can readily down-convert power from either the lower or upper r.f. sideband, or both simultaneously, into the i.f. passband, a filter such as that illustrated in Figure 2.3-6 must be inserted in front of the mixer to define which sidebands enter the mixer.

Because the signal power generally loses strength in the mixing process, the gain G_c associated with this conversion is generally less than unity, and we speak of its reciprocal, the *conversion loss* L_c of the mixer $L_c = 1/G_c = S_2(\text{i.f.})/S_1(\text{r.f., SSB})$. This definition for L_c and G_c explicitly applies to the *single-sideband* (SSB) case where only one of the input sidebands is of interest. We further define the *mixer noise temperature ratio* $t_R = N_2(\text{i.f.})/kT_o$. These definitions for conversion loss and noise temperature ratio for the mixer (often used by manufactures) can be introduced into the standard definition for noise figure to yield:

$$F_{mixer} \triangleq \frac{S_1/N_1}{S_2/N_2} = \frac{S_1/kT_o}{(S_1/L_c)/t_r kT_o} = L_c t_r \tag{2.3.19}$$

Expression (2.3.19) can then be combined with the cascaded noise figure formula (2.3.16) to yield the *superheterodyne noise figure* of the combined mixer and i.f. amplifier:

$$F_{mixer+i.f.\ amp.} = F_{mixer} + \frac{F_{i.f.}-1}{G_{mixer}} = L_c t_r + L_c (F_{if} - 1) = L_c (F_{i.f.} + t_r - 1) \tag{2.3.20}$$

In practice this combined noise figure is typically ~2-8 corresponding to ~3-9 dB, where typical values of t_r are 1.2-1.4 and typical values for L_c are 2-6 dB; larger values are more common for low cost mixers or those operating above ~100 GHz.

The basic receiver architectures for both optical and radio frequency systems have now been introduced. Whether they measure power by squaring the input signal, or correlate the input signal by multiplying it by a reference signal and integrating, they all begin with a bandwidth-limiting element followed directly by the multiplier, or by an intervening amplifier or superheterodyne (frequency-translation) stage. In many cases, particularly in communication systems, either the signal or its square or its cross-correlation with a reference signal are further processed before being presented as output. These complexities are discussed in later chapters.

2.3.4 Multiport networks

Before considering these signal processing complexities, it is useful to review the physical constraints imposed on systems operating on multiple signal streams. Physical linear passive *multiport networks* conserve power and usually exhibit reciprocity $\left(\underline{\underline{Z}}_{ij} = \underline{\underline{Z}}_{ji}\right)$ between ports. Examples of optical multiport networks include *beamsplitters*, *prisms*, *diffraction gratings*, and nonlinear mixers (frequency translators). Common radiofrequency examples include *directional couplers*, *magic tee's*, and mixers. Examples of multiport networks are illustrated in Figure 2.3-9. Orthogonality between ports is ofter useful. For example, the magic tee illustrated in the figure incorporates special metallic protuberances at the junction that provide orthogonality between ports 3 and 4 over much of an octave; ports 1 and 2 are orthogonal at all frequencies because of their physical orthogonality.

Although many linear passive N-port devices are readily characterized by their scattering matrix $\underline{\underline{\underline{S}}}$, insight is sometimes required to identify the ports correctly. For example, a lossy mixer might be characterized as a lossless passive network with one port representing all lossy elements, which are then placed external to the network model. Mixer models typically have such an external resistive component as well as ports at each frequency of interest, which may include all harmonics present, one harmonic per port. Note that the mixer in Figure 2.3-9 omits the L.O. port altogether since no signal of interest enters or leaves by it; L.O. noise is included in the noise port. Waveguide junctions may have one port for each waveguide, but may also have one port for each propagating mode in each waveguide. The nonpropagating modes are each

reactive, and these reactances can generally be lumped together inside a lossless passive N-port network representation. The basic form of an *N-port network* is illustrated in Figure 2.3-8.

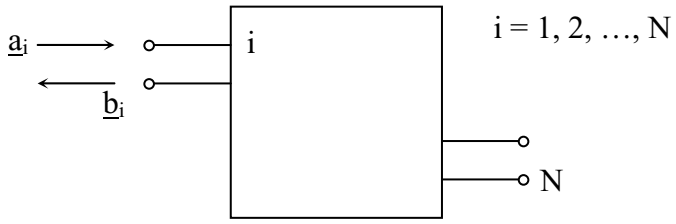

Figure 2.3-8 Linear passive N-port network.

We define the exchangeable power at any port, which can never correspond to more than one mode, as $|\underline{a}_i|^2$ traveling toward port i from the external network, and $|\underline{b}_i|^2$ emerging from port i, where the units typically are watts or watts/Hertz. Both $\underline{a}_i$ and $\underline{b}_i$ are complex quantities, where the phase reference can be defined in various ways. One common definition is:

$$\underline{a}_i = \underline{V}_+ \sqrt{Y_o/2} = \left(\underline{V} + Z_o \underline{I}\right)\Big/\sqrt{8Z_o},\ b_i = \left(\underline{V} - Z_o \underline{I}\right)\Big/\sqrt{8Z_o} \tag{2.3.21}$$

where the phase references for $\underline{V}$ and $\underline{I}$ depend only on the location of reference planes associated with each port, and $Z_o = 1/Y_o$ is the characteristic impedance of the i_{th} transmission line. Other linear combinations of voltage and current can lead to alternate but equally acceptable definitions for $\underline{a}_i$ and $\underline{b}_i$.

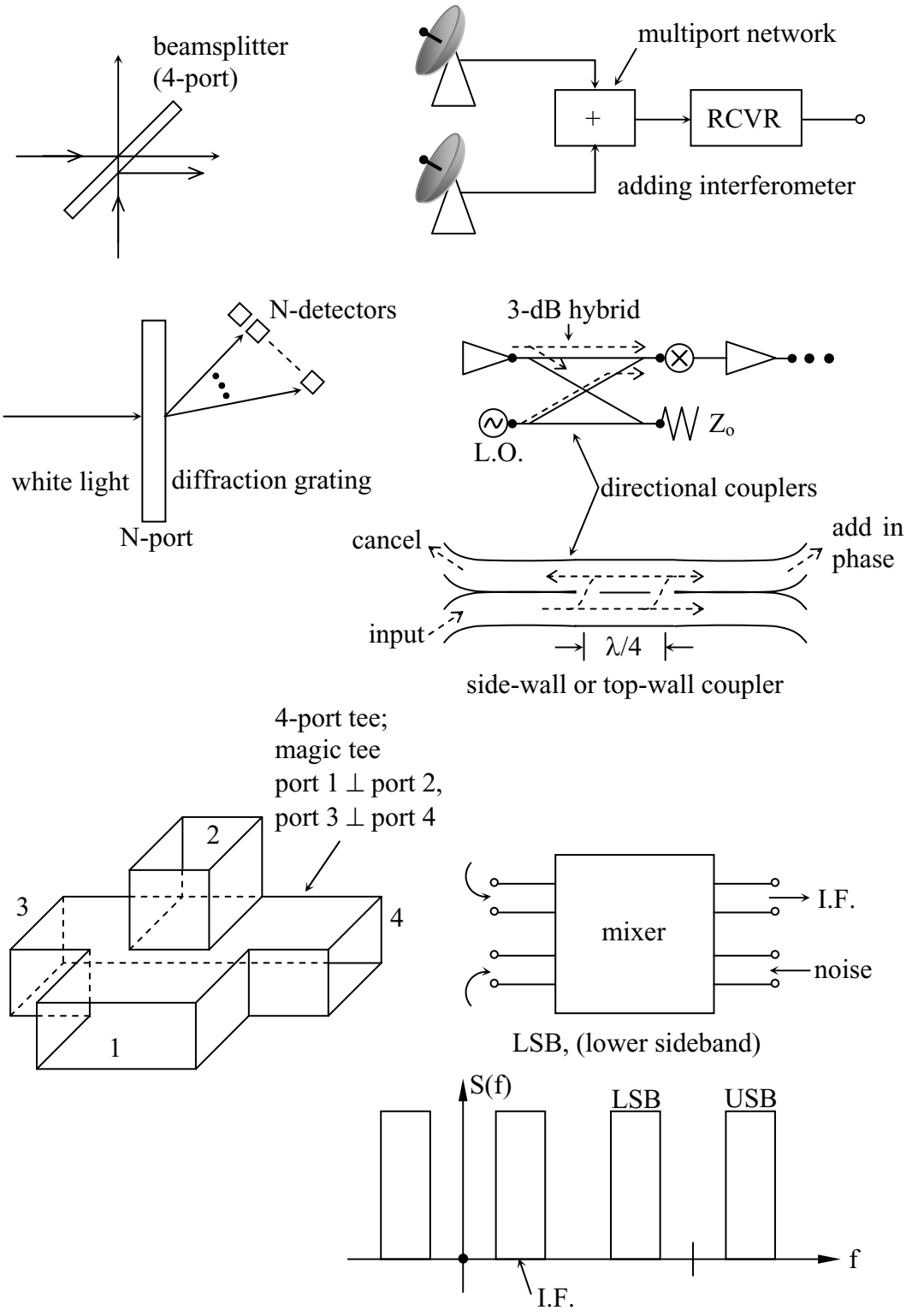

Figure 2.3-9 Examples of multiport networks.

The *scattering matrix* $\underline{\underline{S}}$ relates incoming and outgoing waves $\underline{a}_i$ and $\underline{b}_i$:

$$\underline{\bar{b}} = \underline{\underline{S}}\,\underline{\bar{a}} \tag{2.3.22}$$

It is important to note that the scattering matrix $\underline{\underline{S}}$ is defined only when the N-port network is embedded in a larger network that connects to all ports, thus $\underline{\underline{S}}$ is explicitly dependent on that network. The net power entering port i is therefore $|\underline{a}_i|^2 - |\underline{b}_i|^2$

Between any two ports in an N-port network we may define various types of gain, just as we have for two-port networks. For example we may define the "*transducer gain*" between terminals k and j as:

$$G_{kj} \triangleq |\underline{b}_k|^2 \Big/ |\underline{a}_j|^2 = |\underline{\bar{S}}_{kj}|^2 = P_{Dk}/P_{Aj} \tag{2.3.23}$$

where P_{Dk}/P_{Aj} is the ratio of power delivered to the network from the kth terminal to the power available from the jth terminal of the network.

In general, however, the only gain of interest is the exchangeable gain, which is equivalent to available gain if there are no negative resistances, as defined in (2.3.8) for 2-port networks, and here for N-port networks:

$$G_{E_{k_j}} \triangleq \frac{P_{E_k}}{P_{E_j}} \tag{2.3.24}$$

The externally arriving exchangeable power P_{Ej} is $|\underline{a}_j|^2$ by definition. Evaluation of the power available from the N-port network is more difficult because $|b_k|^2$ is the actual power exiting the N-port network, which has been reduced by any impedance mismatch between the N-port network and the external circuit at port k. To find the exchangeable power at the output we must therefore multiply by a number typically greater than unity which reflects the loss due to any potential mismatch at port k. This coupling coefficient at port k can be evaluated simply, however, by expressing it in terms of both $\underline{a}_k$ and $\underline{b}_k$. This coupling coefficient is:

$$\frac{|\underline{a}_k|^2 - |\underline{b}_k|^2}{|\underline{a}_k|^2} = 1 - |\underline{S}_{kk}|^2 \tag{2.3.25}$$

therefore:

$$P_{E_k} = |\underline{b}_k|^2 \Big/ \left(1 - |\underline{S}_{kk}|^2\right) \tag{2.3.26}$$

Combining (2.3.24) and (2.3.26) yields the desired expression for exchangeable gain between ports k and j:

$$G_{E_{kj}} = \frac{|\underline{S}_{kj}|^2}{1 - |\underline{S}_{kk}|^2} \tag{2.3.27}$$

We can readily impose the constraint of losslessness and passivity on an N-port network by requiring:

$$\sum_{i=1}^{N} |\underline{a}_i|^2 = \sum_{i=1}^{N} |\underline{b}_i|^2 \tag{2.3.28}$$

Reciprocity can be imposed by requiring:

$$\underline{\underline{S}} = \underline{\underline{S}}^t \tag{2.3.29}$$

The mathematical simplicity of the scattering matrix permits efficient derivations of the basic constraints imposed upon N-port networks by the properties of passivity, losslessness, and reciprocity.

For example, we can readily determine whether an ideal power combiner exists. Such a three-port device would be perfectly matched at its two input ports, which would be decoupled one from another, and would transfer the sum of both the input power streams to a single power stream emerging from the output at port 3. These desired properties can be expressed using the scattering matrix as:

$$\underline{\underline{S}} = \begin{bmatrix} 0 & 0 & \underline{\alpha} \\ 0 & 0 & \underline{\alpha} \\ \underline{\alpha} & \underline{\alpha} & \underline{\beta} \end{bmatrix} \tag{2.3.30}$$

where we have assumed symmetry so that $\underline{\alpha}$ accounts for four entries rather than two. That is, the scattering element connecting ports 1 and 3 is the same as that connecting ports 2 and 3, by symmetry. The upper two diagonal elements of $\underline{\underline{S}}$ are zero because they are matched, and the remaining zeros reflect the fact that ports 1 and 2 are not coupled. Reciprocity is imposed in (2.3.30) by the symmetry in $\underline{\underline{S}}$ which exists about the diagonal.

The remaining scattering elements can be further constrained requiring losslessness, using:

$$\underline{\bar{a}}^{t^*} \bullet \underline{\bar{a}} = \underline{\bar{b}}^{t^*} \bullet \underline{\bar{b}} \tag{2.3.31}$$

which follows from (2.3.28). The right-hand side of (2.3.31) can be expressed in terms of $\underline{\bar{a}}$ using (2.3.22), yielding:

$$\underline{\bar{b}}^{t^*} \bullet \underline{\bar{b}} = \left(\underline{\underline{S}}\,\underline{\bar{a}}\right)^{t^*} \bullet \left(\underline{\underline{S}} \bullet \underline{\bar{a}}\right) = \underline{\bar{a}}^{t^*}\,\underline{\underline{S}}^{t^*}\,\underline{\underline{S}}\,\underline{\bar{a}} \tag{2.3.32}$$

To satisfy both (2.3.31) and (2.3.32) simultaneously, it follows that:

$$\underline{\underline{S}}^{t^*}\,\underline{\underline{S}} = \underline{\underline{I}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.3.33}$$

Whether the assumed and quite general form of the scattering matrix expressed in Equation 2.3.30 satisfies (2.3.33) can be readily tested:

$$\underline{\underline{S}}^{t^*}\,\underline{\underline{S}} = \begin{bmatrix} |\underline{\alpha}|^2 & |\underline{\alpha}|^2 & \text{TBD} \\ |\underline{\alpha}|^2 & |\underline{\alpha}|^2 & \text{TBD} \\ \text{TBD} & \text{TBD} & 2|\underline{\alpha}|^2 + |\underline{\beta}|^2 \end{bmatrix} \tag{2.3.34}$$

where the entries labeled by symbol TBD (to be determined) are irrelevant because it is impossible for $|\underline{\alpha}|^2$ to be unity as required for the upper two diagonal elements by (2.3.33), while also satisfying the requirement that the third diagonal element equal unity too. The simple conclusion is that ideal matched symmetric two-input passive power combiners are impossible to build. This example can be repeated with a similar result without the requirement for mechanical symmetry, in which case (2.3.30) would introduce $\underline{\alpha}$, $\underline{\theta}$, and $\underline{\gamma}$ as three unknown scattering elements.

A more complex example involves the question of whether a lossless passive three-port reciprocal network can be matched at all ports simultaneously. The requirement for three matched ports is expressed by requiring zeros on the diagonal of the scattering matrix:

$$\underline{\underline{S}} = \begin{bmatrix} 0 & \underline{\alpha} & \underline{\beta} \\ \underline{\alpha} & 0 & \underline{\gamma} \\ \underline{\beta} & \underline{\gamma} & 0 \end{bmatrix} \tag{2.3.35}$$

Reciprocity is imposed by the diagonal symmetry of the matrix, and passive losslessness is imposed by substituting (2.3.35) into (2.3.33):

$$\underline{\underline{S}}^{t^*}\,\underline{\underline{S}} = \begin{bmatrix} |\alpha|^2 + |\beta|^2 & \beta^*\gamma & \alpha^*\gamma \\ \beta\gamma^* & |\alpha|^2 + |\gamma|^2 & \alpha^*\beta \\ \alpha\gamma^* & \alpha\beta^* & |\beta|^2 + |\gamma|^2 \end{bmatrix} \stackrel{?}{=} \underline{\underline{I}} \tag{2.3.36}$$

The fact that all off-diagonal elements must be zero implies that at least two of the three elements α, β, and γ must be zero, and therefore at least one of the diagonal elements must also be zero, violating (2.3.33) and demonstrating that it is not possible to match all three ports of a linear passive reciprocal network simultaneously. If we violate reciprocity, which can be done by incorporating magnetized ferrites in such a device, then we can show that all three ports can be matched simultaneously, as is commonly done with *three-port circulators* characterized by scattering matrices such as:

$$\underline{\underline{S}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \tag{2.3.37}$$

Such a three-port circulator passes all the power from port 1 into port 2, all the power from port 2 into port 3, and all the power from port 3 into port 1; switchable circulators permit the magnetic field direction to be reversed together with the direction of power flow. It is easily shown that (2.3.37) satisfies (2.3.33).

A still more complicated example of practical importance is the lossless passive reciprocal symmetric four-port device for which ports 1 and 2 are isolated, as are ports 3 and 4. Examples of this are microwave devices called *directional couplers* which can be fabricated by bonding two waveguides together with one or more holes connecting them so as to provide the desired

isolation at the frequencies of interest, (see Figure 2.3-8). Another example is the *optical beam splitter* where an input beam is reflected to one side and also passes straight through, defining three ports, while the return beam is partially deflected to the other side, defining the fourth port. The scattering matrix for this case (see Figure 2.3-8) can be represented as:

$$\underline{\underline{S}} = \begin{bmatrix} 0 & 0 & \underline{\alpha} & \underline{\beta} \\ 0 & 0 & \underline{\beta} & \underline{\alpha} \\ \underline{\alpha} & \underline{\beta} & 0 & 0 \\ \underline{\beta} & \underline{\alpha} & 0 & 0 \end{bmatrix} \tag{2.3.38}$$

That all four ports are matched is imposed by the zeros on the diagonals of (2.3.38), and that the pairs of ports 1,2 and 3, 4 are each isolated is imposed by the other zeros in (2.3.38). Reciprocity is imposed by the diagonal symmetry of the matrix, and mechanical symmetry about the axis separating ports 1 and 3 from ports 2 and 4 is imposed by doubling the appearances of $\underline{\alpha}$ and $\underline{\beta}$ in the scattering matrix. Imposing power conservation (2.3.33) on (2.3.38) results in $|\underline{\alpha}|^2 + |\underline{\beta}|^2 = 1$ and $\underline{\alpha}^*\underline{\beta} + \underline{\beta}^*\underline{\alpha} = 0$. Although $|\underline{\alpha}|^2 = |\underline{\beta}|^2$ is not necessary to satisfy these two results, it is necessary for $\underline{\alpha}\underline{\beta}^* = re^{\pm j\pi/2}$, where the constant r is any real number between zero and one. In this case our constraints resulted in a very specific relative phase shift (π radians) between paths $\underline{\alpha}$ and $\underline{\beta}$ for this four-port network.

2.3.5 Mixers and their noise figures

We may now use this N-port representation to characterize mixers further. When representing real systems by N-port networks, the first step is to define carefully each of the ports of interest. These issues are well represented by the *square-law detector circuit* shown in Figure 2.3-10.

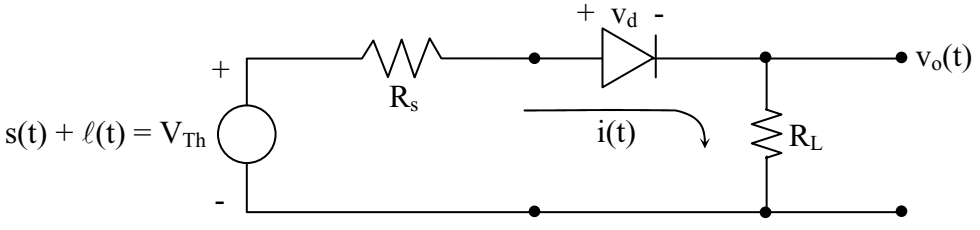

Figure 2.3-10 Square-law detector circuit.

The Thevenin voltage source V_{Th} is a superposition of the signal and local oscillator voltages, $s(t)+\ell(t)$, and it drives current i(t) through the diode and the source and load resistors R_s and R_L; the output voltage $v_o(t)$ appears across the load resistor. The local oscillator signal $\ell(t)$ is proportional to $\sin\omega_o t$ where $\omega_o = 2\pi f_o$. The voltage across the diode v_d is related nonlinearly to the current i(t) by the *diode characteristic*, illustrated in Figure 2.3-11.

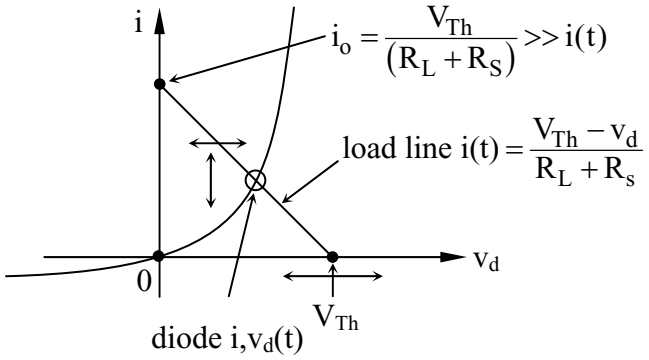

Figure 2.3-11 Diode characteristic, load line, and operating point $v_d(t)$.

Two constraints are imposed upon the relationship between the diode voltage and current; one constraint is imposed by the diode characteristic $i \tilde{\propto} v_d^2$, the other is imposed by the rest of the circuit, which is two resistors in series with the Thevenin voltage source, and is represented by the *load line* equation:

$$i(t)=(V_{Th} - v_d)/(R_L + R_S) \tag{2.3.39}$$

Deviations from an ideal square-law relationship between V_{Th} and i arises partly because $i(v_d)$ is more nearly exponential and the intersection between the load line and the diode characteristic moves in a somewhat nonideal manner. Nonetheless the output voltage can be often approximated to high accuracy by a simple low-order polynomial dominated by the square-law term:

$$v_o(t) = iR_L = \left(\tilde{\propto} v_d^2 R_L\right) = k_o + k_1 v_d + k_2 v_d^2 + k_3 v_d^3 + \tag{2.3.40}$$

The voltage $v_d(t)$ driving the detector is dominated by local oscillator frequency ω_o and the signal ω_s:

$$v_d(t) \cong v_\ell \sin\omega_o t + v_s \sin\omega_s t + \text{small higher-order terms} \tag{2.3.41}$$

The spectral content of the detector output $\underline{V}_o(f)$ can be quite rich because it contains most harmonics of both local oscillator and signal, and all possible sums and differences thereof, as suggested in Figure 2.3-12.

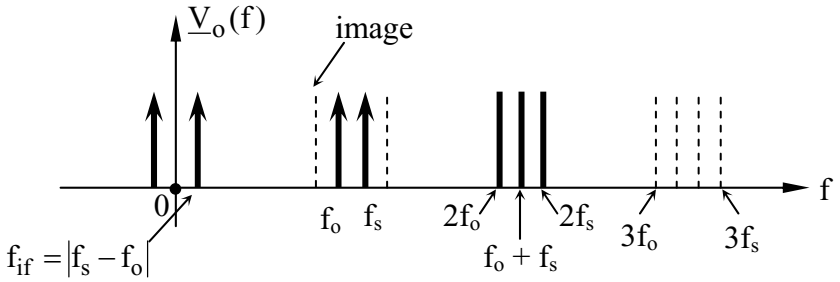

Figure 2.3-12 Spectral content of mixer output.

For a nearly ideal square-law detector the only components with significant power are the local oscillator and signal components at f_o and f_s, and the much lower intermediate frequency f_{if}. It is important that the diode characteristics and associated circuitry suppress higher frequencies because their formation of sums and differences can contribute frequencies that also fall inside the passband of the intermediate frequency amplifier, contributing potentially serious coherent interference.

Figure 2.3-12 also shows the image frequency, which is positioned symmetrically about the local oscillator frequency f_o, opposite signal frequency f_s. Unless a filter preceding the mixer eliminates power entering the mixer at the image frequency, the image will mix with the local oscillator to produce contributions at f_{if}.

The *mixer four-port network model*, illustrated in Figure 2.3-13, can characterize the behavior of most mixers.

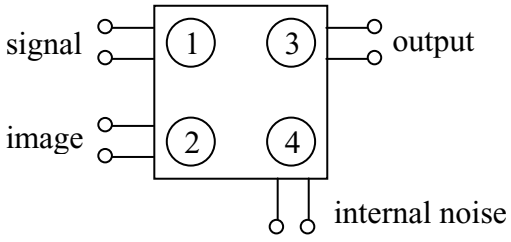

Figure 2.3-13 Four-port network model for a mixer.

The *single-sideband receiver noise temperature* T_R designated T_{SSB}, equals $(F_{SSB}-1)T_o$, where, in terms of scattering matrix elements:

$$F_{SSB} = \frac{S_1/N_1}{S_4/N_4} = \frac{S_1/kT_o}{\left[\dfrac{S_1|S_{41}|^2}{1-|S_{44}|^2}\right] \Big/ \left[\dfrac{kT_o|S_{41}|^2 + kT_2|S_{42}|^2 + kT_3|S_{43}|^2}{1-|S_{44}|^2}\right]}$$

$$= 1 + \frac{T_2}{T_o}\frac{|S_{42}|^2}{|S_{41}|^2} + \frac{T_3|S_{43}|^2}{T_o|S_{41}|^2} = 1 + \frac{T_{SSB}}{T_o} \tag{2.3.42}$$

Therefore,

$$T_{SSB} = T_2\frac{|S_{42}|^2}{|S_{41}|^2} + T_3\frac{|S_{43}|^2}{|S_{41}|^2} = T_o\left(L_c t_r - 1\right) \tag{2.3.43}$$

Note that in (2.3.42) the input noise N_1 is kT_o, the output signal S_4 equals the input signal S_1 times the gain connecting ports 1 and 4, and the output noise N_4 has contributions from port 1, which by definition is terminated with T_o = 290K, plus contributions from the image port (T_2) and the internal resistive port (T_3).

When measuring power spectral density over broad bandwidths it is not uncommon to regard both the signal and image passbands as inputs, which slightly alters the expressions for noise figure and noise temperature. Although it is often suggested that the *single-sideband noise figure* F_{SSB} is twice the *double-sideband noise figure* F_{DSB}, that is true only under certain circumstances. To evaluate F_{DSB} we again use (2.3.42), noting that the signal at the output S_4 is:

$$S_4 = kT_o\left(|S_{41}|^2 + |S_{42}|^2\right)\Big/\left(1-|S_{44}|^2\right) \tag{2.3.44}$$

from which it follows that:

$$T_{DSB} = T_3|S_{43}|^2\Big/\left[|S_{41}|^2 + |S_{42}|^2\right] \tag{2.3.45}$$

If we assume symmetry between signal and image ports, then $|S_{41}|^2 = |S_{42}|^2$ and $T_o = T_1 = T_2$, and:

$$F_{DSB} = 1 + \frac{1}{2}\frac{T_3}{T_o}\frac{|S_{43}|^2}{|S_{41}|^2} \tag{2.3.46}$$

$$F_{SSB} = 2 + \frac{T_3}{T_o}\frac{|S_{43}|^2}{|S_{41}|^2} = 2F_{DSB} \tag{2.3.47}$$

It is easy to see how these expressions could be extended to account for other signal and noise components.

In general, scattering matrixes for multi-port networks can be determined experimentally by introducing test signals and observing their responses at the other ports or frequencies of interest while the device is embedded in the larger network.

2.3.6 Noise cancellation in mixers

A variety of *noise cancellation methods* exist for reducing noise in the receiver types we have considered thus far. Five examples follow. The first two deal with filtering in single-conversion and multiple-conversion superheterodyne systems. The next two examples deal with cancellation of unwanted signals by dividing the signal into two components and then recombining them so that the signal terms add and the noise terms cancel by virtue of the fact that the noise and signal enter the system differently. The fifth example illustrates how calibration techniques can similarly cancel unwanted effects.

Figure 2.3-14 illustrates how local oscillator noise and unwanted signals from the image sideband can be canceled by filtering.

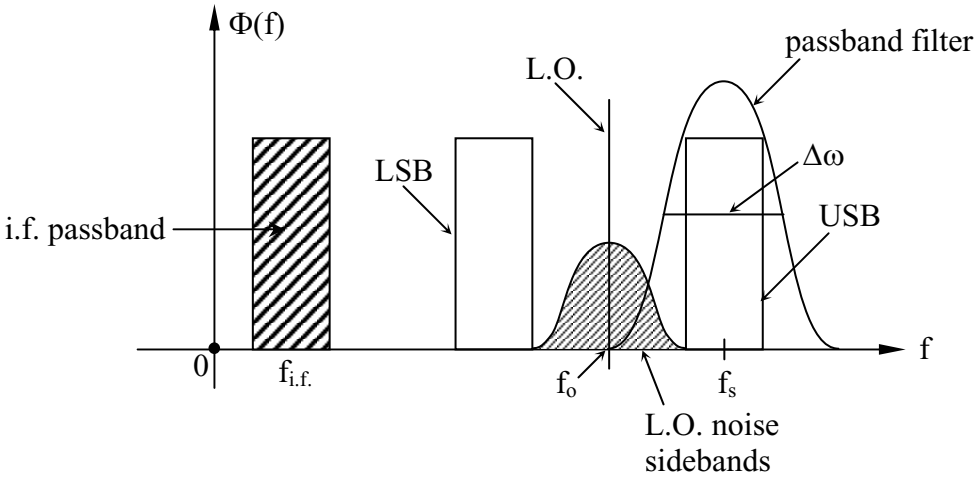

Figure 2.3-14 Signal power spectra for a single-conversion superheterodyne system.

The spectra illustrated in this figure correspond to the case where the signal at f_s resides in the upper sideband (USB). The filter preceding the mixer (bandwidth $\Delta\omega$) is then designed to eliminate signal and noise in the lower sideband (LSB) so that they do not get translated into the

low-frequency i.f. passband centered at $f_{i.f.}$. The figure also illustrates how any noise sidebands introduced by the local oscillator may mix with the local oscillator so as to translate into the i.f. passband, contributing unwanted noise. Most of this noise is eliminated for the case illustrated in the figure by the sharp i.f. passband filter. The illustration suggests how the high-frequency tails of the local oscillator noise may extend sufficiently far to fall inside the passband in any event. Placing a sharp filter on the output of the local oscillator before it enters the mixer is the only easy remedy here. For example, if Q = 100, then the local oscillator should be separated from the signal sideband by at least $2f_o/Q$. Filters implemented with RLC circuits or transmission line equivalents exhibit resonator Q's which typically range from ~50 for RLC circuits to values of ~1000 or more for microwave resonators or even 10^4-10^6 for superconducting, surface acoustic wave, or crystal resonators. Coupling such resonators with active elements and feedback loops can raise Q's even higher.

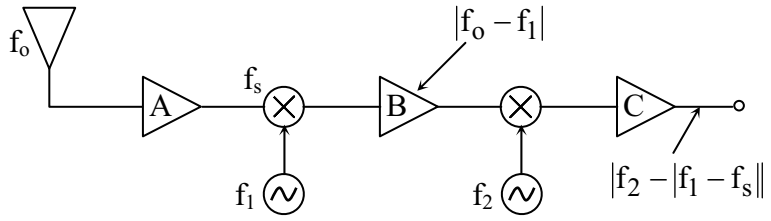

Figure 2.3-15 Dual down-conversion circuit.

Sometimes this frequency separation and its associated intermediate frequency are too great, and thus two down-converters may be cascaded, as suggested in Figure 2.3-15. In general, *multiple conversion* or at least *dual down-conversion* is required when the desired ratio $f_s/f_{i.f.}$ is greater than approximately 1/3 the maximum available Q of the filters. The frequency spectra corresponding to various points in the circuit of Figure 2.3-15 are illustrated in Figure 2.3-16.

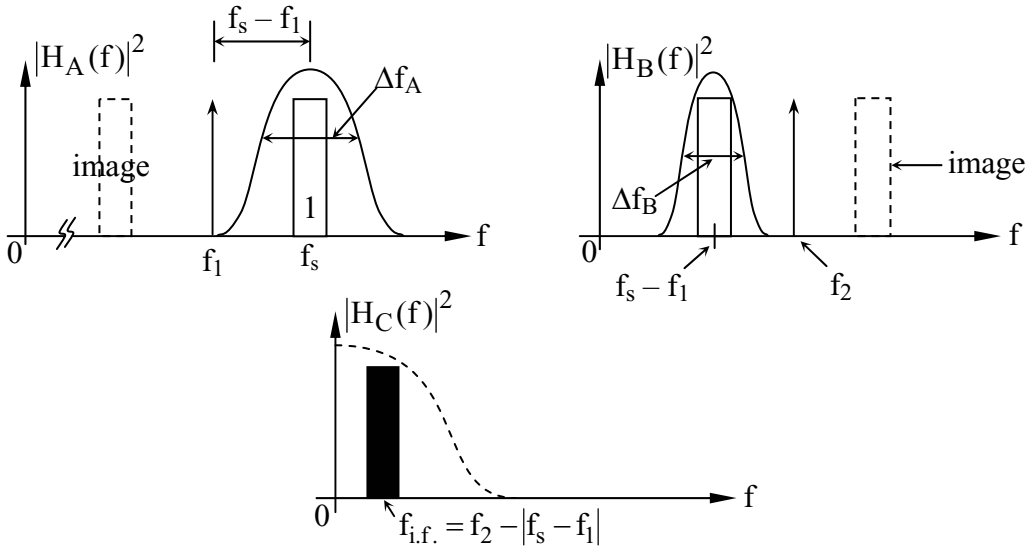

Figure 2.3-16 Power spectra for signals and mixer A, mixer B, and output of a dual down-conversion system.

Consider a 1-GHz wireless phone where each phonecall occupies a fixed 10-kHz bandwidth and these channels have an additional 10-kHz guardband between them; thus the channels are spaced 20 kHz apart. In this case amplifier A of Figure 2.3-15 might have a bandwidth Δf_A adequate to capture hundreds of such adjacent channels, and the first local oscillator f_1 might be separated from the band center f_s by three percent of 1 GHz, or 30 MHz. The bandwidth of the filter which defines the passband of amplifier A can therefore be approximately 30 MHz, which can be realized with a filter Q = 30. In practice, multi-pole filters are often used with $Q > 50$, if such circuit elements are affordable. A filter with comparable Q might be used to reduce any noise sidebands originating from the local oscillator. Both of these filter requirements enforce a minimum separation between the local oscillator and signal frequency that may be too great relative to the narrow bandwidths required for the filters at the intermediate frequencies because $f_{i.f.}$ is too high. In this case the need for narrower output bandwidths can be accommodated by a second mixer with its local oscillator at frequency f_1, as suggested in Figure 2.3-16. In this example f_2 lies above the i.f. passband at $f_s - f_1$. The filter centered in the passband of amplifier B at $f_s - f_1$ serves to reject both the noise sidebands of the local oscillator at f_2 and the image band which might otherwise be translated into the passband of the amplifier C; the second $f_{i.f.} = f_2 - |f_s - f_1|$, as suggested in Figure 2.3-16. In this example amplifier A might be centered near 1 GHz, amplifier B might be centered near 10 MHz, and the final amplifier C might be centered near 100 kHz. Then a multipole filter with $Q \cong 100$ could produce a reasonable boxcar filter of 10-kHz width, sufficient to block signals from adjacent channels from entering any particular channel passband, and to allow the rolloff of these channel-definition filters to lie within the nominal 10-kHz guardbands separating adjacent channels. Note that crystal filters

have $Q \tilde{>} 10^5$, so a single-conversion superheterodyne incorporating them could provide these same 10-kHz channels, if desired.

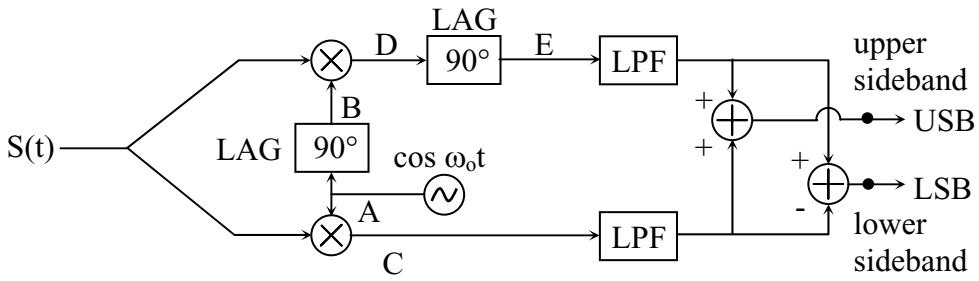

Figure 2.3-17 Superheterodyne sideband cancellation circuit.

The next example illustrates how signal processing rather than filtering can be used to cancel the image sideband while preserving intact the signal sideband. Filters are more difficult to implement when the signals to be separated are narrow compared to their absolute frequencies; fortunately this is the circumstance for which the signal processing technique illustrated in Figure 2.3-17 works best. In this system the signal s(t) enters a matched 3-port network which divides the power equally, conveying it to two separate mixers driven by the same local oscillator, but with a 90-degree lag in one arm for which the output experiences a second 90-degree lag before being amplified in a lowpass filter (LPF) similar to the one at the output of the other mixer. The upper sideband is selected by adding these two filtered outputs, while the lower sideband is selected by subtracting them.

To show that this system works, it is useful to recall two identities:

$$\cos \omega_o t = \frac{e^{j\omega_o t} + e^{-j\omega_o t}}{2} \tag{2.3.48}$$

$$\sin \omega_o t = \frac{e^{j\omega_o t} - e^{-j\omega_o t}}{2j} \tag{2.3.49}$$

These expressions for the cosine and sine functions lead to the graphic representation for local oscillator A driving the lower mixer in Figure 2.3-17, and local oscillator B driving the upper mixer. These spectra are shown in Figure 2.3-18.

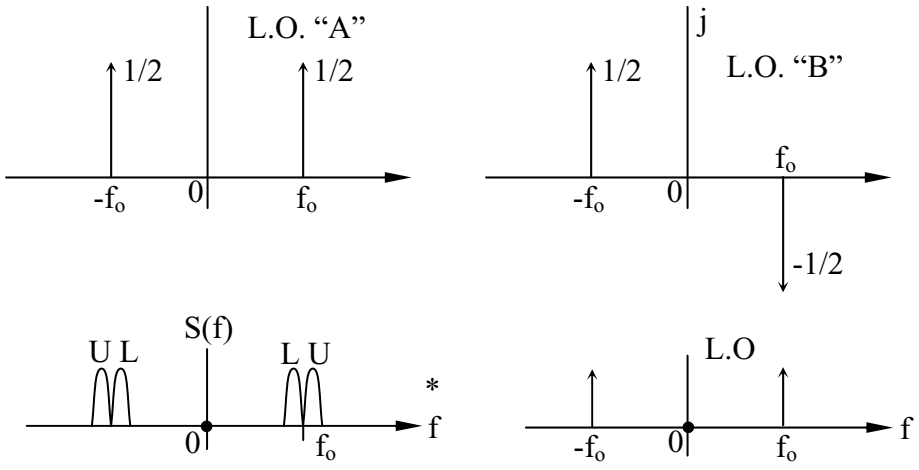

Figure 2.3-18 Local oscillator spectra in a sideband-cancellation circuit.

The double-sided spectrum for the signal S(f) shows the relationships between the upper sideband U, the lower sideband L, and the local oscillator A.

Multiplication of s(t) and local oscillator A to yield signal C of Figure 2.3-17 results in convolving their spectra in the frequency domain, as suggested in Figure 2.3-19. The convolution C of S(f) and local oscillator A is shown in the top left part Figure 2.3-19 where we see the upper and lower sidebands have been superimposed in the same frequency band; our task is to separate them. The spectrum for signal D is the convolution S(f) with the transform of sin $\omega_o t$, also as illustrated in Figure 2.3-19; in this case the i.f. frequencies convey the difference between the upper and lower sidebands, shifted 90 degrees (× j). The signal E emerging from the 90-degree lag inserted after the mixer has the spectrum shown at the bottom left of Figure 2.3-19, where the i.f. frequencies again represent the difference between the upper and lower sidebands, but without any phase shift relative to the output signal C from the other mixer. It is now easy to see how summing signals C and E yield the spectrum shown on the top right of Figure 2.3-19, which is the upper sideband alone. The lower sideband is the difference between the signals C and E. The bandwidth over which this sideband-separation circuit works well is limited by the bandwidth of the 90-degree phase shifters. In general the bandwidths of such systems are less than a few percent for sideband cancellation factors greater than ~20 dB.

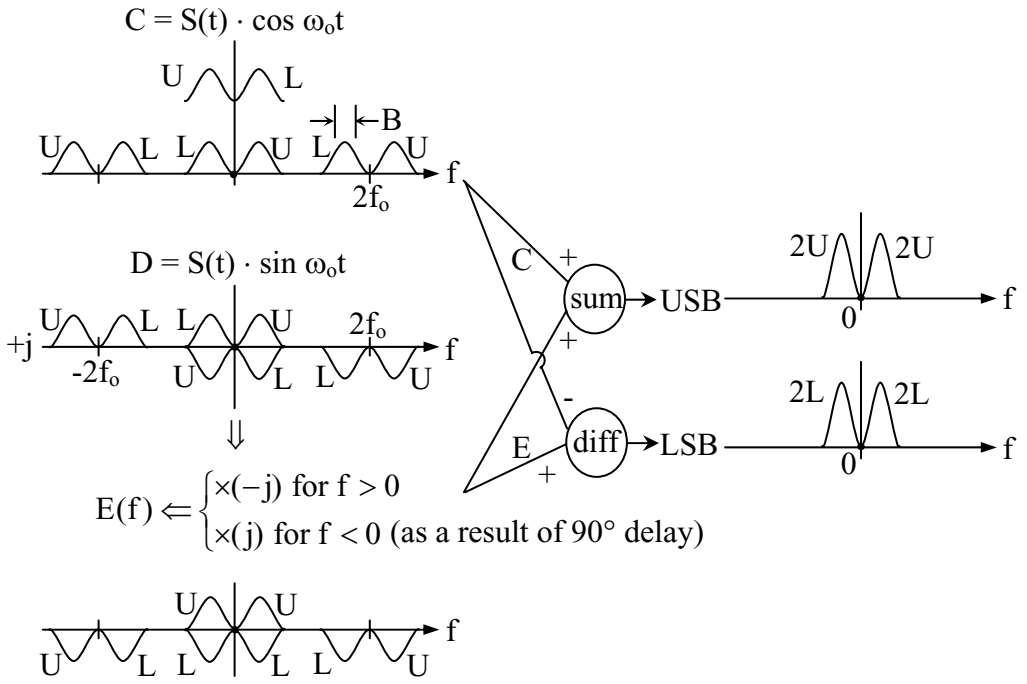

Figure 2.3-19 Intermediate-frequency spectral distribution in a sideband cancellation circuit.

Cancellation is used in mixers not only for selecting sidebands, but also for cancelling noise sidebands of the local oscillator which otherwise might mix with the local oscillator to produce beat frequencies within the i.f. passband. The circuit and waveforms in Figure 2.3-20 illustrate the relationship between diode orientation and i.f. signal phase.

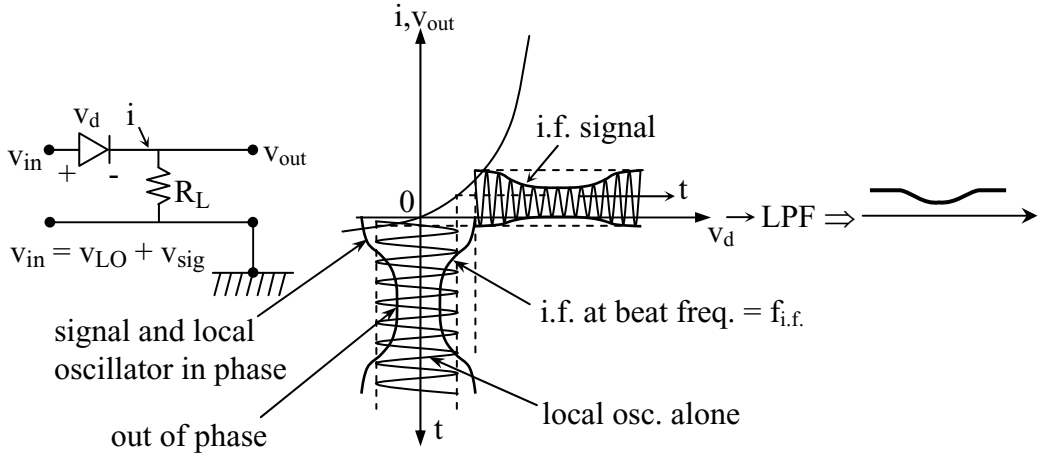

Figure 2.3-20 Mixer circuit and waveforms.

The output voltage v_o is directly proportional to the diode current i flowing through the load resistor R_L. The diode characteristic $i(v_d)$ is plotted with the current i along the vertical axis and the diode voltage v_d along the horizontal axis. The vertical axis is also used to represent time, so that the local oscillator signal can be represented as a sine wave which slowly drifts in-phase and out-of-phase with the signal at the i.f. beat frequency $f_{i..f.}$. The resulting current i(t), as illustrated in the figure with time now on the horizontal axis, is an r.f. sine wave modulated at the intermediate frequency much more strongly for positive voltages than for negative voltages. When this output intermediate frequency voltage is low-pass filtered, the emerging signal closely resembles the positive envelope of the combined local oscillator and signal sinusoids, but with half the peak amplitude.

The signal and local oscillator waveform can be combined in a variety of devices, one of which is the magic-tee, illustrated in Figure 2.3-21.

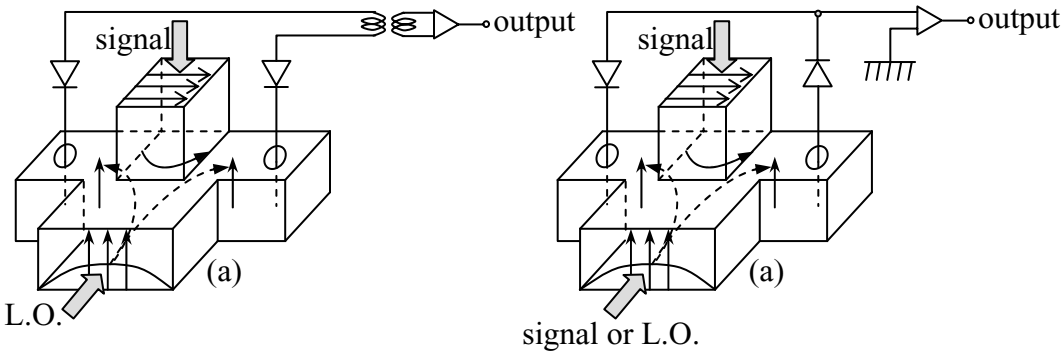

Figure 2.3-21 Mixer configuration using a magic-Tee.

A magic-tee is fabricated by welding four waveguides together as illustrated, incorporating small reactive elements at the junction that, at the design frequency, isolate the two lateral ports from each other and permit all four ports to be matched simultaneously when embedded a matched network. The signal and local oscillator ports are orthogonal by virtue of geometry, and the other two ports are isolated from each other by virtue of junction reactances even though there is generally a clear physical path connecting them.

In Figure 2.3-21a the local oscillator signal enters from the front and splits equally between the two side arms in phase, which can be seen from the illustrated electric field configuration of the dominant mode of the waveguide. The signal enters from the top producing out-of-phase electric fields in each of the two side arms due to the antisymmetry of the junction. If the two diodes have wires extended across the waveguide so they can pick up the propagating electric fields, then the intermediate frequency signal detected by the two diodes will be out-of-phase due to the sign reversal associated with the signal relative to the in-phase local oscillator signal in the two diode wires. This can also be seen by considering the voltage waveforms in Figure 2.3-20. These out-of-phase intermediate frequency signals readily pass through the output transformer to

be amplified subsequently. Any local oscillator noise, however, effects both diodes identically, and therefore any noise contribution at the intermediate frequency cancels in the same transformer. Such cancellation is typically 20-30 dB, where higher values of isolation are difficult to obtain without very precise matching of the diode and circuit characteristics.

A variation of this configuration is suggested in Figure 2.3-21b where reversed diodes are used, and the intermediate frequencies are added. In this case the local oscillator noise has opposite polarities in the two diode outputs, and such summation cancels it, while permitting the signals to add.

An alternate four-port network used for mixers is illustrated in Figure 2.3-22.

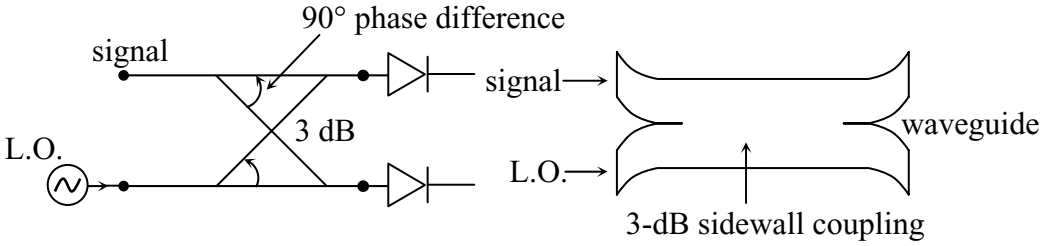

Figure 2.3-22 Mixer employing 3-dB sidewall coupler.

It can be shown (see Section 2.3.4) that a passive, lossless, symmetric, reciprocal 4-port network must exhibit a 90-degree phase difference between signals exiting the two output ports; since this applies to signals that enter through either the signal or the local-oscillator port, the signal and local oscillator each experience a cumulative 180-degree phase reversal relative to one another at the two diodes, yielding performance similar to that of the magic-tee.

Efficiency of mixers can be better understood and improved by also considering their time-domain performance. Consider the idealized diode illustrated in Figure 2.3-23.

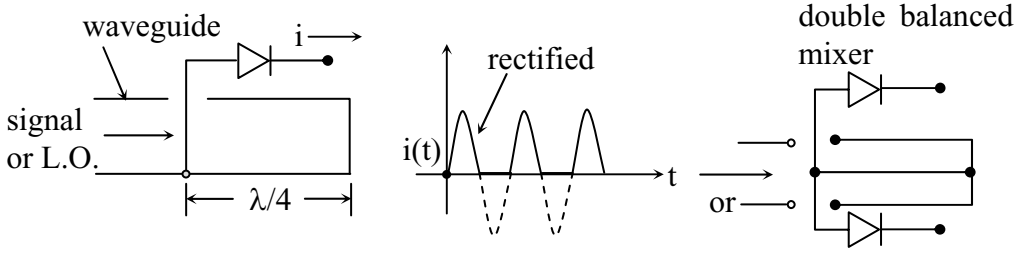

Figure 2.3-23 Time-domain currents and single- and double-balanced mixers.

Over short time periods the superposition of the signal and local oscillator signals resembles a sinewave, as illustrated, which is picked up and rectified by the diode as it feeds the i.f. amplifier. During those half-wave intervals when the diode is open-circuited, the

electromagnetic wave can pass largely unimpeded down the waveguide to be reflected by the short circuit approximately one-quarter wavelength away. This time delay and the phase reversal at the wall result in the return wave doubling the current through the diode without changing its strongly pulsed character (half-wave rectification). In contrast, the illustrated double-balanced mixer draws current through the upper diode for one half of the r.f. cycle, while the other diode draws current during the other half. Since less reliance is placed on the quarter-wave delay, the intrinsic bandwidth may be broader.

2.3.7 Calibration techniques

Cancellation techniques often contribute importantly to accurate calibration of receiving systems. These issues are well illustrated by a case study involving early measurements of the 2.7 K isotropic cosmic background radiation as measured from a mountaintop. The challenge is not only to calibrate the relationships between antenna temperature and voltage, as suggested in Figure 2.3-24, but also to measure the atmospheric contribution.

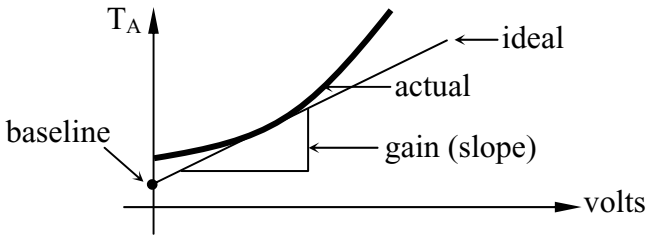

Figure 2.3-24 Calibration curve for non-ideal receiver.

The relationship between antenna temperature and receiver output voltage generally departs slightly from the desired ideal linear relationship, which is usually characterized by a baseline offset and a gain. These can be calibrated as suggested in Figure 2.3-25.

The receiver in Figure 2.3-25 consists of a mixer driven by a local oscillator for which the frequency is measured by a tunable resonant absorbing frequency meter located in front of a crystal detector. The signal enters the mixer from a ferrite circulator used as an isolator. This 3-port device transports power entering from one port to the next clockwise output port, for all three ports. Thus any local oscillator leakage passes first to a matched load and thus is attenuated 20-30 dB before entering the Dicke switch, which in this diagram is a 4-port switchable circulator. The switchable circulator can move energy to adjacent ports either clockwise or counterclockwise, depending on the direction of the magnetizing current. Thus for this circuit the receiver looks alternately at an antenna pointed at cold space, or a switch that views either the main antenna or the liquid helium calibration load through identical horn antennas. Terminating the fourth port of this circulator in a cold sky view reduces the amplitude of any thermal noise it might indirectly introduce to the Dicke switch or that might be reflected from an imperfect calibration switch.

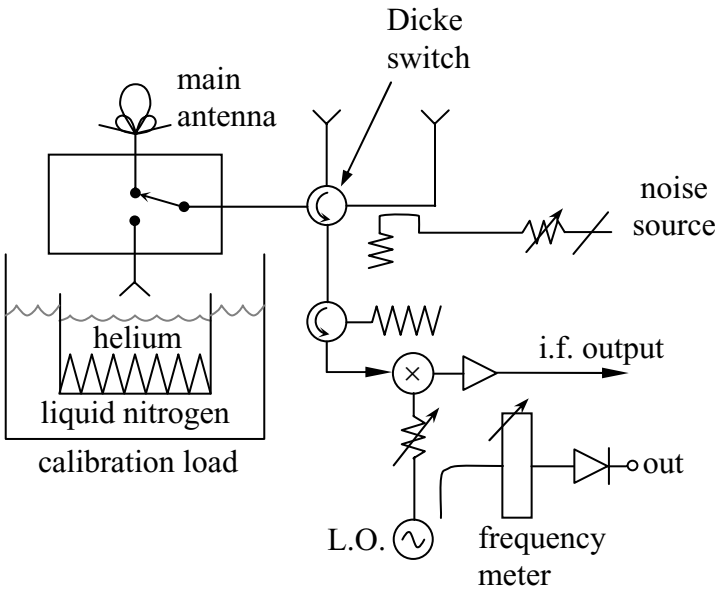

Figure 2.3-25 Radio receiver calibration circuit.

The baseline offset can be calibrated by switching between the main antenna and a calibrated blackbody load, in this case a liquid helium-bathed set of absorbing pyramids inside an evacuated dewar floating in liquid nitrogen. Reflections from the surface of the liquid helium are reduced by tilting the dewar so that the received polarization crosses the interface near Brewster's angle. The dewar is further capped by a metallic reflector that returns any reflections to the cooled load.

The gain is calibrated by a noise generator that can be turned on or off while switched into the receiver input. The noise generator can be calibrated in turn by comparing its incremental contributions to the receiver output with contributions from the main antenna as it alternately views the cold sky (near 6K) and a calibrated blackbody near ambient temperature (measured by a thermometer).

The symmetry of the Dicke switch in Figure 2.3-25 can be tested by connecting all three input ports to identical sky antennas observing termperatures far below ambient and noting any differences. The symmetry of the main helium/sky switch can be tested by reflecting the duplicate antennas 90° skyward by a special antenna calibration reflector assembly. The symmetry of this assembly can be tested by flipping it relative to the horizontally pointing main antennas.

A particularly important test ensures no strong signals radiated by the local oscillator or mixer will reflect back into the receiver differently depending on the position of the main calibration switch. A tunable short circuit replaces the main calibration switch while being viewed by the Dicke switch. If this tunable short, as it moves through a full half wavelength,

induces no change in the radiometer output, then any stray escaping local oscillator signal could not produce a systematic calibration error if the calibration switch and load VSWR is a function of switch position. Such calibration-switch dependent L.O. reflections commonly interfere coherently with the original L.O. signal to modulate the detected power as a function of calibration switch position, introducing troublesome calibration errors that must be remedied.

A certain amount of energy is also emitted by the partially cooled metallic sidewalls of the matched load submerged in helium. These losses can be measured by removing the calibration load from the liquid helium and letting the system view the sky through the same pipe, but with the loads at the bottom removed. By observing the loss as the length of this pipe is doubled, some sense of wall emission can be obtained; values less than ~0.2K are anticipated from room temperature walls if the original antenna beamwidth viewing the helium is $<\sim 20^{o}$.

Most difficult to calibrate is the contribution of atmospheric emission, which is due principally to water vapor below ~40 GHz. Figure 2.3-26 illustrates how a receiver on a mountain top might view zenith and then perhaps 60 degrees from zenith where the optical pathlength through the atmosphere, and therefore the atmospheric radiance, is approximately doubled for nearly transparent atmospheres. At frequencies near 8-mm wavelength the mountaintop atmospheric absorption coefficient is perhaps one percent, corresponding to zenith emission of ~3K. If this doubles at 60^{o} zenith angle, then the zenith emission can be estimated and subtracted from the observed value to yield the background sky brightness temperature, near 2.7K.

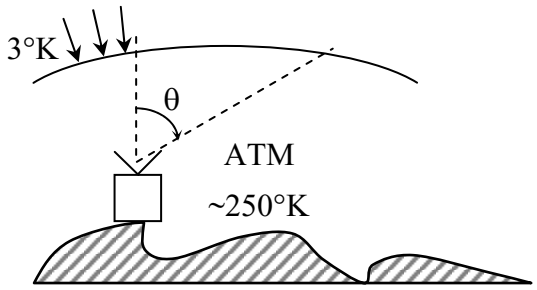

Figure 2.3-26 Atmospheric emission measurement deduced from an elevation scan of radiance.

2.4 OPTICAL AND INFRARED DETECTION

2.4.1 Photoelectric detectors

The preceding discussion in Section 2.3 applies in the Rayleigh-Jeans (radio) limit where the photon energy hf is much less than the thermal energy kT. Next we consider the opposite optical limit where the photon energy hf is much greater than kT, and the infrared case where they are comparable. The discussion below begins with a review of common types of photon detectors.

Photodetectors sense the movement of electrons ionized by photons impacting metals in vacuum or semiconductors. Prior to the advent of semiconductors, sensitive photodetectors utilized the photoelectric effect wherein an arriving photon elevates an electron sufficiently far above the metallic Fermi level that it can climb over or tunnel through the potential barrier which exists within ~2nm of the surface of the photo-emitter. The height of this barrier is called the *work function* of the metal and is approximately 4-5 volts for most common metals, and drops to 1.95, 2.1, and 2.3 volts for cesium, rubidium, and lithium, respectively. These potentials are sensitive to surface contamination and microstructure, typically dropping when the surface has sharp points which can concentrate electric fields, and increasing if there are additional insulting barriers. This work function approximates the energy associated with a charge being attracted by its mirror image as it approaches the surface, where the closest approach approximates the diameter of the outermost electron orbitals.

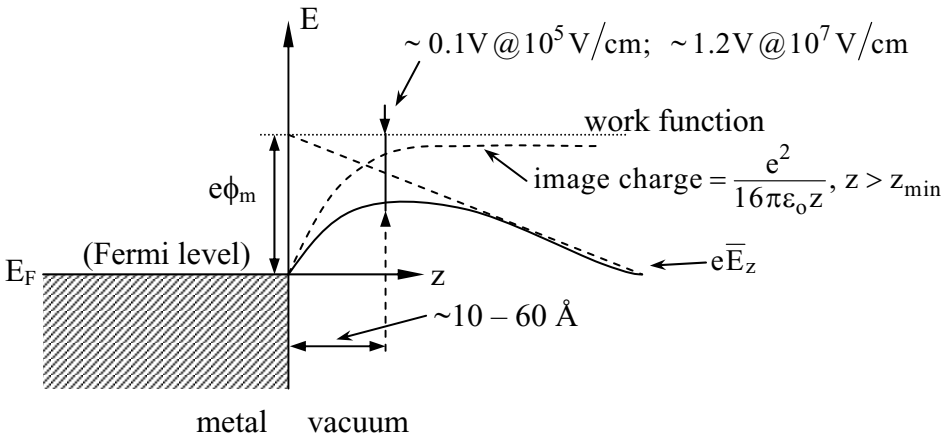

Figure 2.4-1 Photoelectric potentials near a metallic surface and vacuum.

In order for photoemission to occur the photon energy hf should be approximately greater than the metallic work function $\phi_{m\,e}$. If this potential is one volt, then hf equals e joules or one electron volt (1 e.v.). The cutoff wavelength for such a detector is:

$$\lambda_{c.o.} = c \Big/ f_{c.o.} = \frac{hc}{e\phi_m} \cong 0.6 - 0.7\mu m \text{ for cesium} \tag{2.4.1}$$

Such a photo emittor can be excited from the front or, if it is sufficiently thin, from the back, as suggested in Figure 2.4-2.

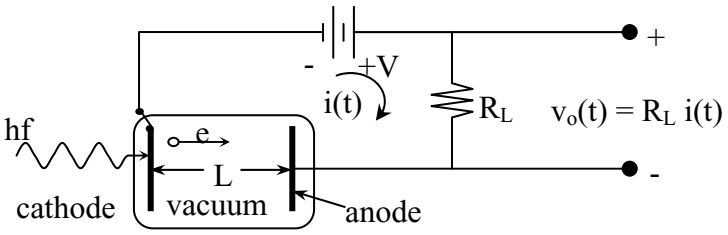

Figure 2.4-2 Phototube detector circuit.

In this circuit a typical photon may release an electron with a probability equal to the quantum efficiency η, which is typically less than 30 percent for germanium or silicon cathodes. The released electron then accelerates linearly toward the anode where it is intercepted. The current pulse in the anode circuit associated with this moving electron increases linearly with time as the image charge on the anode increases, and ends when the electron impacts the positively charged anode, neutralizing that image charge. This current i can be characterized by the electron charge e times the electron velocity u divided by the cathode-anode gap L. That is, the anode current i(t) is approximately:

$$i(t) = eu/L = eat/L \cong e^2 Vt \Big/ m_e L^2 \tag{2.4.2}$$

where the acceleration of the electron a equals the force on the electron eV/L divided by the electron mass m_e. If photon arrival times are rare compared to the transit time of the electrons, then the anode current is a Poisson-distributed series of triangle waves for which the integrated current is one electron charge. A practical problem with such detectors is that these weak currents and the voltages $v_o(t) = R_L i(t)$ are typically small compared to the Johnson noise produced by the load resistor R_L.

Photomultiplier tubes (PMT's) were developed to overcome this Johnson noise. They have the form illustrated in Figure 2.4-3. The electrons emitted by the photons impacting such a PMT are accelerated to perhaps 100 volts, sufficient for the impacting electron to eject several secondary electrons from the first anode, called a dynode, which is positioned so that an additional ~100-volts potential exists between it and a second dynode, where again each of the secondary electrons can now produce several secondaries of its own. A typical PMT might have 7-13 dynodes collectively producing an electron gain of $\sim 10^4$-10^7. Now the current pulse and the associated voltage spike is normally much larger than the Johnson noise from R_L so that individual incoming photons can be detected with high reliability, provided they produce the

initial photoelectron. Such phototubes typically detect 1000 or more events per second originating from comic rays, thermal emission, and environmental radioactivity.

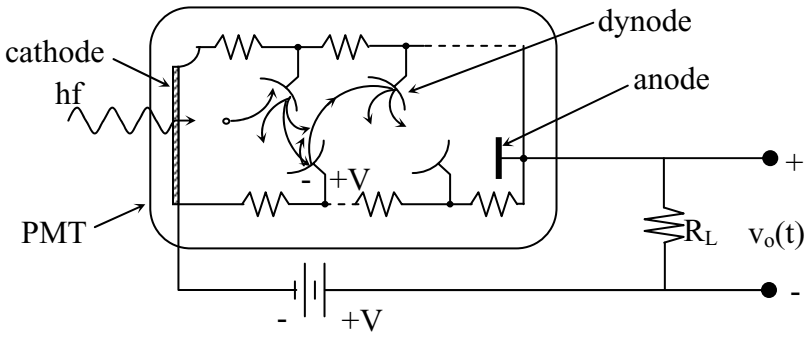

Figure 2.4-3 Photomultiplier tube circuit.

Those electrons originally emitted from other than the first dynode typically produce current pulses which are several times smaller than those produced normally, so that electronic circuitry can reject them. Typical photomultiplier tubes can produce counting rates of ~10MHz-1GHz or more before the pulses begin to overlap.

2.4.2 Semiconductor photodiodes

Similar photon emission processes occur inside semiconductor diodes near the junctions between p-type and n-type semiconductors. Consider the semiconductor energy diagram shown in Figure 2.4-4. The figure also illustrates the density of electron states in the conduction and valence bands, both of which increase towards the gap as illustrated by the boldface crosshatching. The vertical axis represents electron energy and the horizontal axis represents position in a direction parallel to the pn junction. The *valence band* of energy is associated with electrons occupying bound positions whereas any higher energy electrons located in the *conduction band* are free to move; the conduction band is empty at zero temperature.

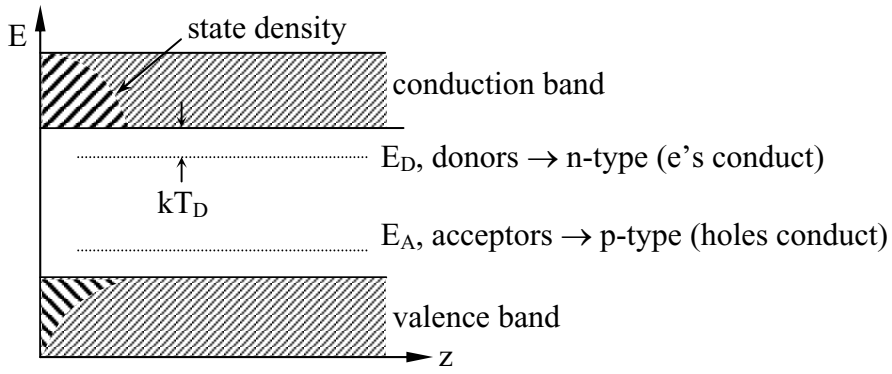

Figure 2.4-4 Electron energy diagram for doped semiconductors.

Although electrons can move between the valence and conduction bands, they occupy positions in the energy gap between the two bands only when bound to an impurity *donor atom* at energy level E_D or an acceptor at energy E_A. Lattice defects can also trap electrons. Electrons bounds to donors at energy E_D can readily be ionized at low temperatures or by photons or phonons with (low) energies greater than kT_D (see Figure 2.4-4). When ionized to the conduction band these electrons become the dominant carriers. Such a semiconductor doped with donors is called an *n-type semiconductor* because negative carriers dominate conduction. Alternatively, a semiconductor may be doped predominately with acceptors having energy levels E_A located sufficiently close to the valence band that thermal or photon energy can readily excite electrons from the valence band across the small energy gap where they become bound to the *acceptor atoms* which sparsely populate the semiconductor. As these electrons leave the valence band they also leave a vacancy behind called a "*hole*" which can readily be filled by adjacent electrons, causing the hole to move. We regard such conduction in the valence band as being dominated by positive carriers called "holes", even though such a hole moves as a vacancy filled by physically translating electrons one atom at a time.

Such semiconductors have a nominal electron energy associated with the top of the electron population. At high temperatures this *Fermi level* (analogous to sea level) lies in the middle of the band gap. At room temperatures the Fermi level lies close to E_A for p-type semiconductors and close to E_D for n-type semiconductors. Thus when an open-circuited p-n junction is formed between p-type and n-type semiconductors, the Fermi level of all three must align as suggested in Figure 2.4-5.

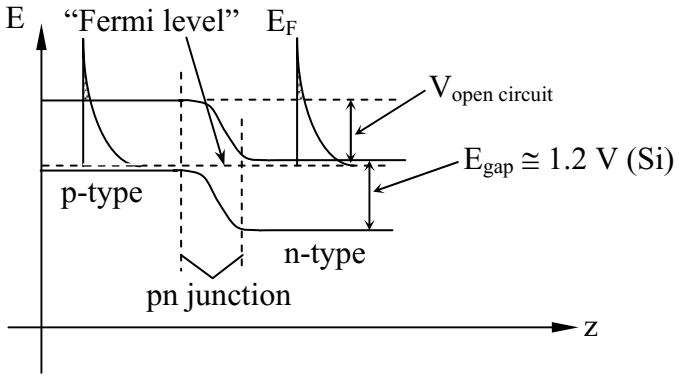

Figure 2.4-5 Electron energy diagram for a pn junction

Note how the Fermi level remains at constant energy across the junction whereas the energies associated with the bottom of the conduction band and top of the valence band are at non-constant energies. In the case of silicon the gap energy E_{gap} is approximately 1.2 volts. When the voltages are as shown, the Boltzmann distribution of electrons above the Fermi level in the n-type semiconductor has an energy tail sufficient to inject electrons into the p-type region; this tail is essentially identical to the Boltzmann tail associated with electrons which might be excited from the Fermi level in the p-type semiconductor. That is, there is equilibrium between electrons excited on each side of the junction and propagating toward the other side; these two currents balance and are in equilibrium if the junction is open circuited.

If we bias the pn junction positively or negatively, current will flow. Figure 2.4-6 suggests how the current in a forward-biased diode can increase exponentially with the bias voltage. The Boltzmann distribution of thermally excited electron energies in the n-type side of the junction has an exponentially increasing number of electrons with energies above the bottom edge of the conduction band of the p-type semiconductor, and they therefore can move freely to the left of the diagram, corresponding to a diode current flowing from the p-type to n-type side of the junction. When the diode is forward biased the i-v characteristic illustrated in Figure 2.4-6 results. The thermally excited electrons in the n-type semiconductor are trapped and cannot cross the junction. Only the few thermally excited electrons rising above the low-lying Fermi level in the p-type semiconductor are able to flow; these may be supplemented by current associated with holes left behind by excited electrons in the n-type semiconductor. This reverse current flow is largely independent of reverse bias and approximates I_o, as suggested by the i-v characteristic of Figure 2.4-6.

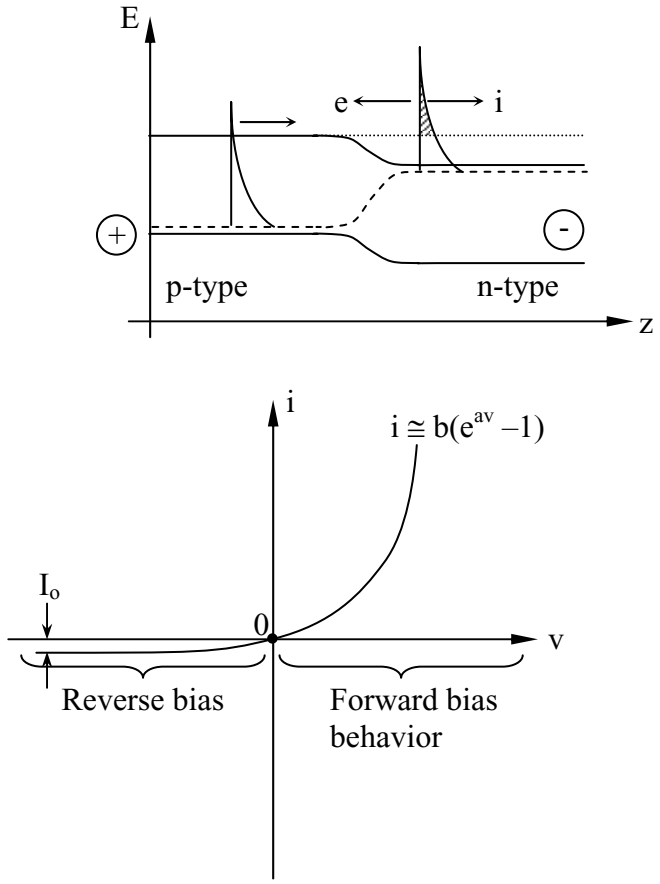

Figure 2.4-6 Energy diagram for a forward-biased diode.

The behavior when the diode is negatively biased is suggested in Figures 2.4-6 and 2.4-7. The thermally excited electrons in the n-type semiconductor are trapped and cannot cross the junction. Only the few thermally excited electrons rising above the low-lying Fermi level in the p-type semiconductor are able to flow; these may be supplemented by current associated with holes left behind by excited electrons in the n-type semiconductor. This reverse current flow is largely independent of reverse bias and approximates I_o, as suggested by the i-v characteristic of Figure 2.4-6.

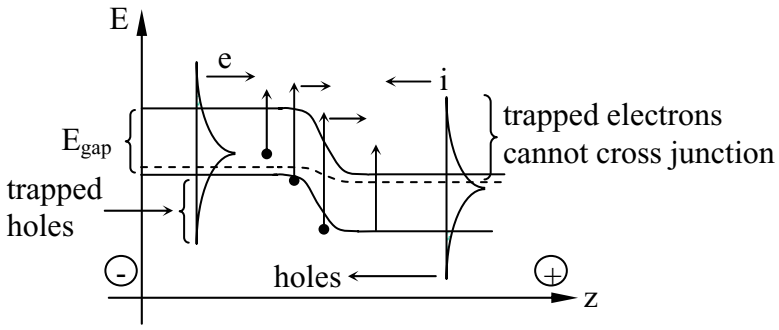

Figure 2.4-7 Electron energy diagram for a reverse-biased diode.

Such p-n junctions can function as photodetectors when photons excite electrons into the conduction band so they can exit from the negative terminal of the reverse-biased diode. Thus the reverse current in the diode increases linearly with photo-excitation, and nonlinearly with diode temperature T. To minimize the number of thermally excited electrons, sensitive photodiodes are typically cooled to reduce their dark current to acceptable levels. If the photon energy is sufficiently high, i.e. $hf > E_{gap}$, then the quantum efficiency can be ~0.8-0.95 conduction electrons produced per photon.

A typical photodiode circuit is shown in Figure 2.4-8.

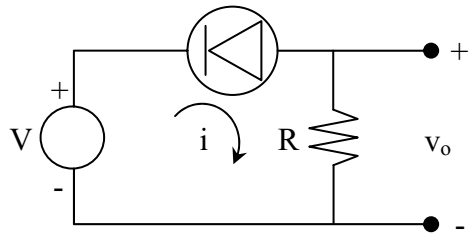

Figure 2.4-8 Photodiode circuit.

The output voltage of this circuit is:

$$v_o = i_{dark} + \frac{\eta P_s}{h\nu} eR \tag{2.4.3}$$

where the photon power $P_s = nhf$, and where n is the number of photons incident on the diode junction per second.

2.4.3 Avalanche photodiodes

A more sensitive detector can be obtained by reverse biasing such photodiodes more strongly into the "avalanche" region, creating an avalanche photodiode (APD). A typical i-v characteristic for an avalanche photodiode is shown in Figure 2.4-9. When biased in the avalanche region an APD exhibits markedly increased current for small increases in reverse-bias voltage v.

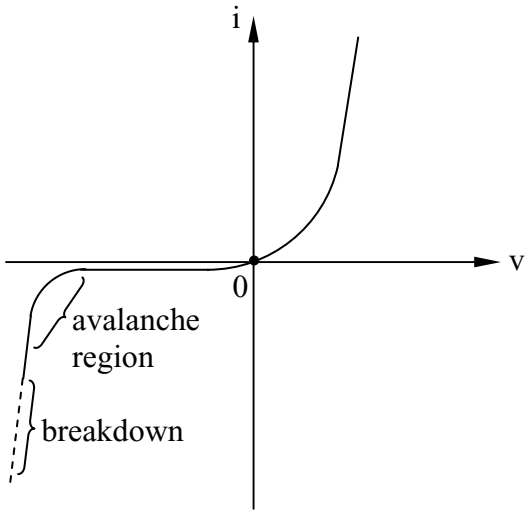

Figure 2.4-9 Avalanche photodiode i-v characteristic.

This is because each excited electron has an energy sufficiently above the conduction band edge in the n-type semiconductor that it can be accelerated by the induced electric field so as to knock one or more additional electrons out of the lattice so they too join the avalanche, as suggested in Figure 2.4-10. The figure shows how a photon with energy hf photo-excites an electron from the Fermi level into the conduction band where it eventually loses energy by collisionally exciting another electron across the same gap. They become a pair which can further collisionally excite more electrons until none have energy sufficient for further excititations.

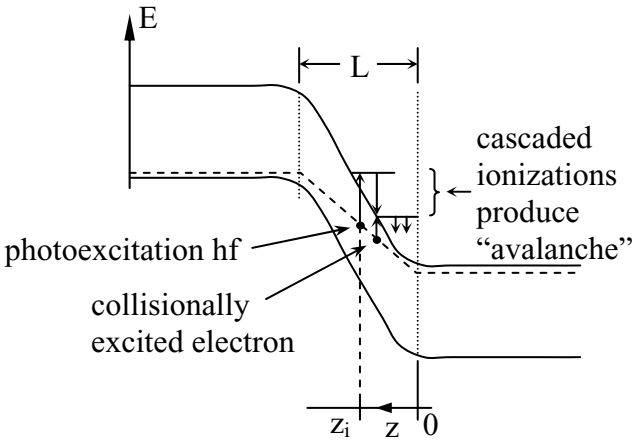

Figure 2.4-10 Electron energy diagram for a backed-biased avalanche photodiode

The number of electrons produced in an avalanche from a single photoexcitation is limited by the exponential growth process associated with the avalanche. The i^{th} photon has gain $g_i \cong e^{g_o z_i}$, where g_o is a constant and z_i is approximately the distance into the junction where the excitation occurs and is therefore proportional to the electrical potential and energy extractable from the collisionally excited electron, as suggested in Figure 2.4-10. A photoexcited electron gains energy as it moves toward the positive terminal and then collides with a bound electron, releasing it; this process repeats yielding exponential growth in the current pulse.

The average gain G associated with the single photoexcitation is therefore:

$$G = E\left[e^{g_o z_i}\right] \cong \frac{1}{L}\int_o^L e^{g_o z} dz = \frac{1}{g_o L}\left(e^{g_o L} - 1\right) \tag{2.4.4}$$

where L is the junction thickness and the maximum value of z_i. Although this equation only approximates the avalanche process, it is roughly correct and has some intuitive value. Later we shall find use for the expected value of g^2, which is:

$$E\left[g^2\right] = E\left[e^{2g_o z_i}\right] \cong \frac{1}{2g_o}\left[e^{2g_o L} - 1\right] \tag{2.4.5}$$

$E[g^2]$ appears later in an expression for current fluctuations in APD's associated with the effective detector noise. The nonuniformity of the current produced in each photoexcitation pulse is suggested in Figure 2.4-11, where the pulse heights tend to be uniformly distributed in a logarithmic space.

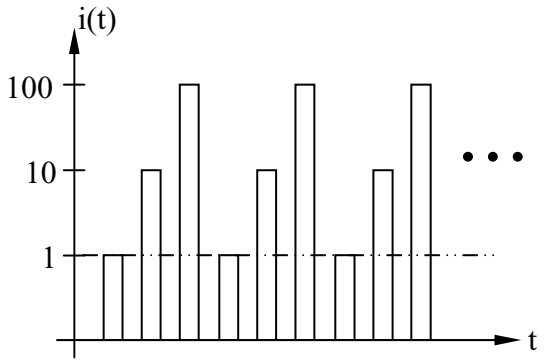

Figure 2.4-11 Poisson-distributed pulses with random amplitudes

Thus:

$$E\left[g^2\right]/G^2 \cong G^x \tag{2.4.6}$$

where $x \cong 0.25$, and values of 0.2-0.5 are typical. Typical values for average gain chosen in practice are ~23 ±6dB.

2.4.4 Photodetector carrier-to-noise ratio

Photodetectors are often usefully characterized by their *carrier-to-noise ratio* (CNR), where:

$$\text{CNR} \triangleq \overline{i_s(t)}^2 \Big/ E[i-\bar{i}]^2 \tag{2.4.7}$$

where $i_s(t)$ is the signal current and i is the total current flowing through the load resistor R_L in the circuit of Figure 2.4-12. CNR is closely related to SNR.

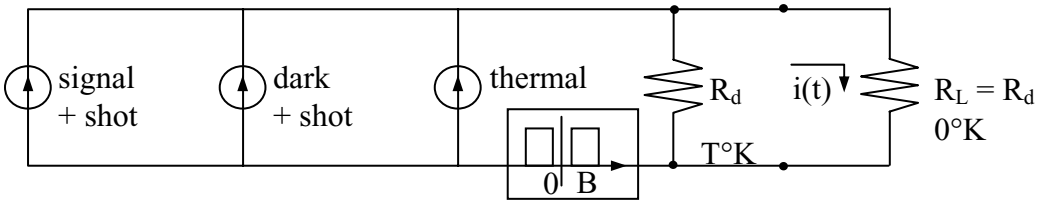

Figure 2.4-12 Photodetector circuit model.

The average signal current $\overline{i_s(t)}$ in (2.4-7) equals one electron charge e per detected photon, where the number of detected photons is the quantum efficiency η times the received signal power P_s divided by the energy per photon hf. For avalanche photodiodes the charge generated per detected photon is eG; in this case:

$$\overline{i_s(t)} = \eta P_s \, eG/hf \tag{2.4.8}$$

The variance of the total current i has contributions from thermal noise contributed by R_d and shot noise associated with the signal and dark-current photons. The variance of the signal and dark current associated with shot noise is:

$$\overline{i_{n\,shot}^2} = 2B(eG)\bar{i}_{S+D} \tag{2.4.9}$$

which follows from (2.1.36). If we assume the load resistor R_L is matched to the detector impedance R_d but contributes no thermal noise of its own (i.e. is at 0^oK), then the thermal noise current i_n divides equally between R_d and R_L. Since the thermal power flow to the load kTB equals $\overline{(i_n/2)^2}R_L$, it follows that:

$$\overline{i_{n\,thermal\ R_L}^2} = 4kTB/R_L \tag{2.4.10}$$

It then follows from (2.4.7-10) that:

$$CNR = \frac{(\eta P_s \, eG/hf)^2}{\underbrace{(2BeG\eta(P_s + P_D))eG/hf}_{i_{n\,shot}^2} + \underbrace{(4kTB/R_L)}_{i_{n\,thermal}^2}} \tag{2.4.11}$$

which applies to photodiodes or phototubes with unity gain, or photomultiplier tubes with gain G > 1. Manipulating (2.4.11) yields:

$$CNR = \frac{\eta P_s/hf\,2B}{1 + [P_D/P_S] + 2k\,Thf/\left[R_L \eta P_s (eG)^2\right]} \tag{2.4.12}$$

Equation 2.4.12 yields the best possible CNR if the temperature T and the dark current P_D both approach zero, or the signal power P_s approaches infinity; this is the quantum limit. In this quantum limit we want large quantum efficiency η and signal power P_s, and small bandwidth B. The option of letting R_L approach zero to produce an infinite denominator in (2.4.12) is not practical because we assumed R_L equals R_D, and any mismatch would be counterproductive. In practice designers choose T so that $P_D < P_s$ and choose G so that $R_L G^2$ is sufficiently large that the thermal noise term in the denominator of (2.4.12) is small compared to unity. Sometimes economics prevents these limits from being achieved.

The CNR for avalanche photodiodes (APDs) has an additional noise contribution from the variable gain of an APD. That is, the shot noise due to poisson arrival times is enhanced by the variable amplitudes of each current impulse, as suggested in Figure 2.4-13.

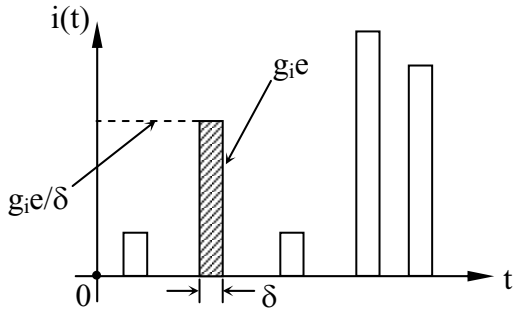

Figure 2.4-13 Avalanche photodiode current idealized (rectangular pulses).

The integral under each impulse equals the charge e of an electron times the instantaneous gain of the APD; a pulse duration of δ as it appears in the figure corresponds to a current $i = g_i e/\delta$. The number of $\bar{n}$ of such impulses per second equals $\eta(P_s + P_D)/hf$. The autocorrelation function for this current, and its corresponding power density spectrum are illustrated in Figure 2.4-14.

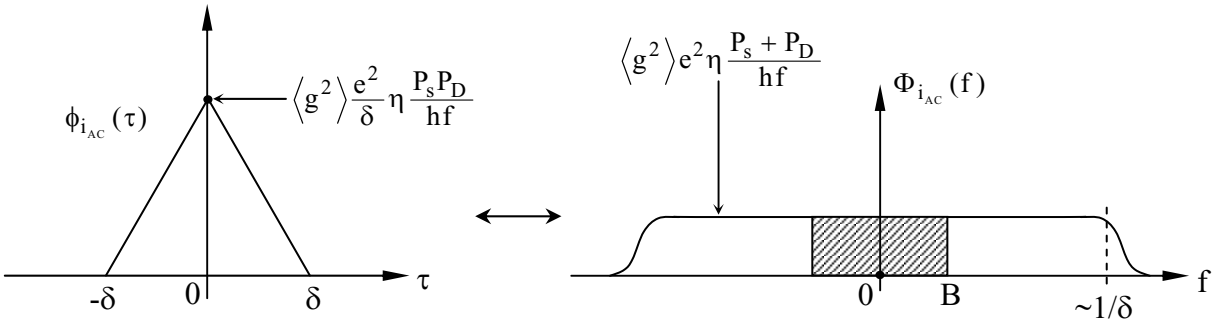

Figure 2.4-14 APD current autocorrelation function and power density spectrum.

The variance σ_i^2 of the current flow from DC to bandwidth B corresponds to the hatched portion of Figure 2.4-14, and is:

$$\sigma_i^2 = \left[(i - \bar{i})^2\right] = \int_{-B}^{B} \Phi_i(f) df = 2B \left\langle g^2 \right\rangle e^2 \eta (P_s + P_D)/hf \qquad (2.4.13)$$

Using (2.4.13) in (2.4.7) we obtain:

$$\mathrm{CNR(APD)} = \frac{(\eta P_s eG/hf)^2}{\underbrace{2B\left\langle g^2\right\rangle e^2\eta(P_s+P_D)/hf}_{\sigma_i^2} + \underbrace{4kTB/R_L}_{i_n^2\,\mathrm{thermal}}} \qquad (2.4.14)$$

$$= \frac{\eta P_s/hf2B}{\frac{\left\langle g^2\right\rangle}{G^2}\left(1+\frac{P_D}{P_s}\right)+\frac{2kThf}{R_L\eta P_s(eG)^2}}$$

It can be shown that:

$$\left\langle g^2\right\rangle \Big/ G^2 \cong G^x \text{ where } x \cong 0.2\text{-}0.5 \qquad (2.4.15)$$

Optimization of CNR typically leads to designs where G^2 is sufficiently large that thermal noise becomes negligible, but still sufficiently small that G^x remains modest; typically $G^x \cong 4$.

2.4.5 Bolometers for infrared detection

Analysis of infrared detectors is somewhat more complex than analysis of "radio" detectors for which hf<<kT and "optical" detectors for which hf>>kT. The most common such detector is a *bolometer*, which focuses the incoming radiation on a heat-sensitive element having a resistance R that depends on the temperature T. For the circuit illustrated in Figure 2.4-15 the incoming signal power Ps heats R(T) thus changing the output voltage v_{out}. Because such detectors are most sensitive at low temperatures, they are thermally coupled to a heat sink at a bath temperature T_b by a thermal conductivity G_t (watts/K). The thermal conductivity of a heat pipe equals $\Delta P_t/\Delta T$, where ΔT is the temperature difference across the heat pipe and ΔP_t is the resulting flow of thermal power through the heat pipe. Typical bath temperatures T_b are the boiling points of helium (4K) and nitrogen (77K).

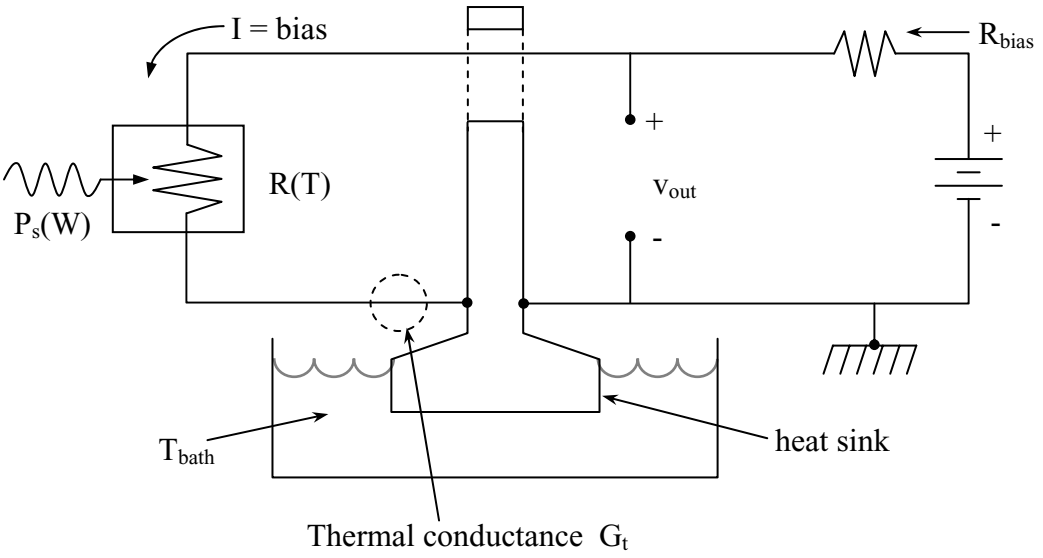

Figure 2.4-15 Bolometer circuit diagram

Although some simple bolometers detect heat by bending a bi-metallic strip or producing motion by heating a gas which expands, sensitive detectors usually employ n-type semiconductors containing donors in the band gap at energy kT_d below the bottom of the conduction band, as suggested in Figure 2.4-16.

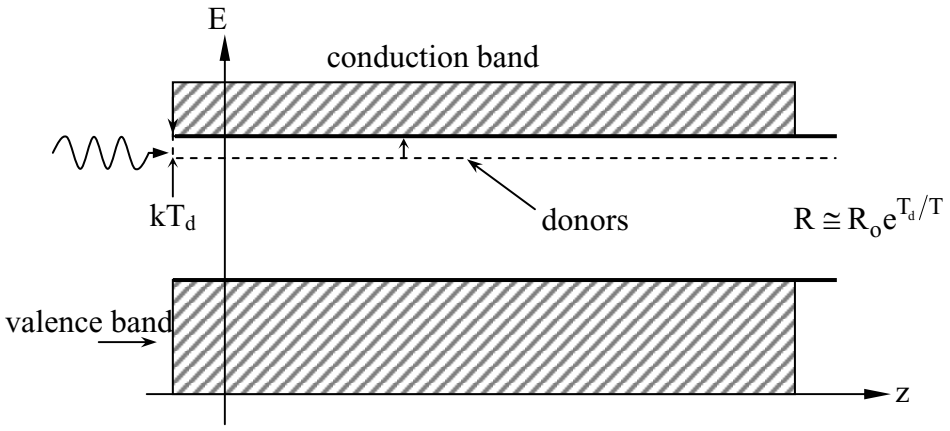

Figure 2.4-16 Bolometer detector energy diagram

Photons with energy greater than kT_d can ionize donor atoms and introduce electrons into the conduction band so as to decrease the detector resistance R where:

$$R \cong R_o e^{T_d/T} \tag{2.4.16}$$

The operating temperature T_b of the detector should be sufficiently small compared to T_d that photon excitation is the dominant process for detector ionization.

Bolometer responsivity S is the incremental sensitivity of the output voltage v_o to changes in the input signal power P_s:

$$S \triangleq \frac{\partial v_o}{\partial P_s} = I\frac{\partial R}{\partial P} \bullet \frac{\partial P}{\partial P_s} \text{ where } P = I^2R + P_s \tag{2.4.17}$$

The total power input to the detector includes ohmic heating I^2R plus the absorbed photon power P_s; differentiating this total power P with a respect to signal power P_s yields:

$$\frac{\partial P}{\partial P_s} = I^2\frac{\partial R}{\partial P}\frac{\partial P}{\partial P_s} + 1 = \left(1 - I^2\frac{\partial R}{\partial P}\right)^{-1} \tag{2.4.18}$$

where R_{bias} is typically much larger than the detector resistance R so that the bias current I can be assumed constant and independent of P_s. Combining (2.4.17) and (2.4.18) yields:

$$S = I\frac{\partial R}{\partial P} \Big/ \left(1 - I^2\frac{\partial R}{\partial P}\right) \text{ where } \frac{\partial R}{\partial P} = \frac{\partial R}{\partial T} \bullet \frac{\partial T}{\partial P} \tag{2.4.19}$$

We note $\frac{\partial T}{\partial P} = 1 \Big/ G_t$ and:

$$\frac{\partial R}{\partial T} = \frac{\partial}{\partial T}\left(R_o e^{T_d/T}\right) \cong \frac{-RT_d}{T^2} \tag{2.4.20}$$

where we have assumed that the energy gap kT_d is small compared to the thermal energy associated with the bath operating temperature T_b of the detector, i.e. $T_d/T_b << 1$. Combining (2.4.19) and (2.4.20) we find the desired expression for bolometer responsivity:

$$S = \frac{-IT_dR}{G_tT^2} \Big/ \left(1 + \frac{I^2T_dR}{G_tT^2}\right) \tag{2.4.21}$$

Equation 2.4.21 suggests that there is an optimum bias current I, since responsivity S approaches zero in the limits where I approaches either zero or infinity.

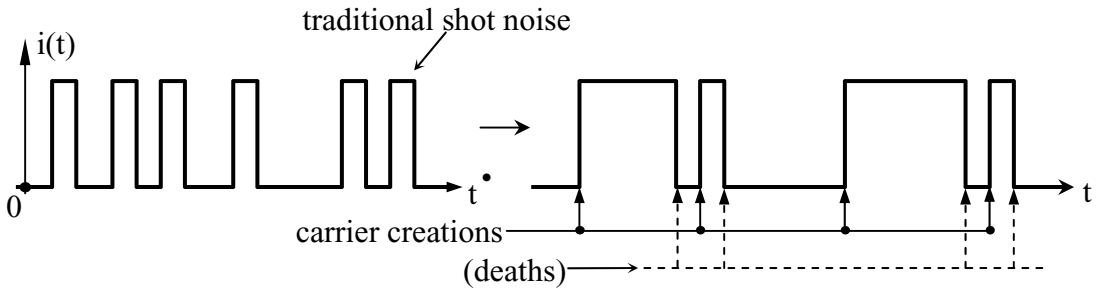

Figure 2.4-17 Waveforms exhibiting both shot noise and random recombinations, or recombination noise

Four different noise processes can contribute to the total bolometer noise. First, thermal noise can be contributed by the resistances in the detector R and in the bias resistor R_b. The photon noise statistics will be a combination of the statistical behavior in the optical and radio limits. In addition, since phonons mediate the heat flux to the thermal bath and are discrete, they will contribute *phonon noise*. There can also be *recombination noise* due to the finite random lifetimes of excited carriers; it generally increases the shot noise associated with carrier creation by less than a factor of two. Figure 2.4-17 suggests how random recombinations in a photoconductor can increase the shot noise associated with current flow through it.

It is customary to characterize noise in infrared detectors in terms of their *noise-equivalent power* (NEP), the dimensions of which are Watts $Hz^{-1/2}$. For example, from (2.1.30) we see that the rms Johnson noise voltage produced by a resistor R at temperature T is :

$$V_J = \sqrt{4kT_bR}\left(\text{v Hz}^{-1/2}\right) \tag{2.4.22}$$

For the bolometer of Figure 2.4-15 where $R_{bias} >> R$, the Johnson noise contribution to v_{out} is dominated by the Johnson noise from the detector resistance R because the voltage divider diminishes the Johnson noise from the bias resistor by a factor of $R/(R+R_B)$. In this limit,

$$NEP_J = V_J \Big/ \left(\frac{\partial V}{\partial P_s}\right) = V_J/S \tag{2.4.23}$$

This expression is directly analogous to that used earlier for the sensitivity of a total-power radiometer:

$$\Delta T_{rms} = v_{o_{rms}} \Big/ \frac{\partial v_o}{\partial T_A} \tag{2.4.24}$$

Although *phonon noise* is difficult to evaluate exactly, simple approximations provide some intuitive understanding. Phonons are associated with acoustic waves propagating through solids

at frequency f, where $hf \cong kT$, as usual. Figure 2.4-18 suggests how the slightly unbalanced phonon fluxes of n_R and n_L phonons per second are moving to the right and to the left, respectively, between the photoconductor chip and the thermal bath through a path with thermal conductivity G_t. The net thermal flux into the bath averages P_+watts and the temperature drop from chip to bath is ΔT.

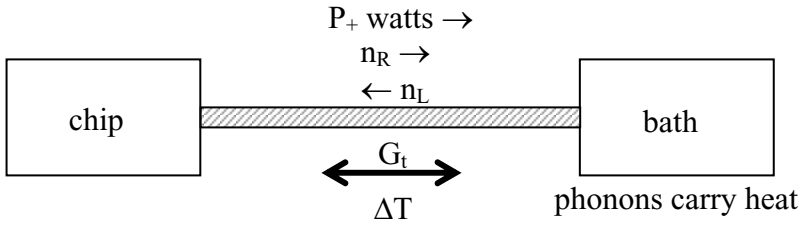

Figure 2.4-18 Phonon thermal transport in a bolometer

The net heat flux P_+ into the bath is:

$$P_+ \cong k\left(n_R T_{chip} - n_L T_{bath}\right) \cong kn\Delta T\left(n = n_R \cong n_L\right) = G_t \Delta T \tag{2.4.25}$$

The random fluctuations in detector temperature associated with the random heat flux exiting the device can approximated for this simple model as:

$$NEP_{G_t} \cong \sigma_{P_+} kT\sqrt{2n} \cong kT\sqrt{2G_t/k} = \sqrt{2kG_t T^2} \tag{2.4.26}$$

which closely approximates the standard answer:

$$NEP_{G_t} \cong \sqrt{4kG_t T^2} \quad \left[W\ Hz^{-1/2}\right] \tag{2.4.27}$$

This expression for phonon noise suggests it can approach zero if the thermal conductivity G_t approaches zero. However, this option would substantially increase the detector temperature T, and therefore there is an optimum conductivity G_t.

To compute the radiation contribution NEP_R to the total detector NEP we first need to compute the variance σ_n^2 of the number n_i of photons in the i^{th} mode of a resonating cavity, where $E_i = hf_i$. The Maxwell-Boltzmann distribution of photons in thermal equilibrium defines the probability p(n) of having n photons with frequency f:

$$p(n) = D\ e^{-n[hf/kT]} \tag{2.4.28}$$

where D is a constant to be determined using the constraint that:

$$\sum_{n=0}^{\infty} p(n) = 1 = \sum_{n=0}^{\infty} D\, e^{-nx} \tag{2.4.29}$$

where x = hf/kT. It is useful to define:

$$S_k(x) \triangleq \sum_{n=0}^{\infty} n^k e^{-nx} \tag{2.4.30}$$

then it follows that:

$$S_{k+1}(x) = \frac{-d}{dx} S_k(x) \tag{2.4.31}$$

$$S_o(x) = 1 \Big/ \left(1 - e^{-x}\right) \tag{2.4.32}$$

and the sum (2.4.29) becomes:

$$\sum_{n=0}^{\infty} D\, e^{-nx} = D \bullet S_o = \sum_{n=0}^{\infty} p(n) = 1 \tag{2.4.33}$$

$$D = 1 \Big/ S_o = 1 - e^{-hf/kT} \tag{2.4.34}$$

The expressions above simplify evaluation of the variance of the number of photons per mode, σ_n^2:

$$\sigma_n^2 \equiv E\left[\left(n - \bar{n}\right)^2\right] = E\left[n^2\right] - \left(E[n]\right)^2 \equiv \overline{n^2} - \bar{n}^2 \tag{2.4.35}$$

where the average number of photons per mode:

$$\bar{n} = \sum_{n=0}^{\infty} np(n) = \sum_{n=0}^{\infty} n \frac{1}{S_o} e^{-nx} = S_1 / S_o = e^{-x} \Big/ \left(1 - e^{-x}\right) \tag{2.4.36}$$

$$\overline{n^2} = \sum_{n=0}^{\infty} n^2 p(n) = S_2/S_o = \frac{S_1}{S_o} + \frac{2e^{-2x}}{\left(1-e^{-x}\right)^2} = \bar{n} + 2\bar{n}^2 \tag{2.4.37}$$

where, applying (2.4.31 to (2.4.32) it follows that:

$$S_1(x) = e^{-x} \Big/ \left(1-e^{-x}\right)^2 \tag{2.4.38}$$

$$S_2(x) = S_1(x) + 2e^{2x} \Big/ \left(1-e^{-x}\right)^3 \tag{2.4.39}$$

Combining equations (2.4.35-39) yields the desired results:

$$\sigma_n^2 = \overline{n^2} - \bar{n}^2 = \bar{n} + \bar{n}^2 \tag{3.4.40}$$

Equation (2.4.40) says, in general and in the infrared regime, that $\sigma_n \cong \left(\bar{n} + \bar{n}^2\right)^{1/2}$. Since $\bar{n} = \left(e^{hf/kT} - 1\right)^{-1}$, it follows that in the optical limit where hf>>kT we have $\bar{n} << 1$ and $\sigma_n \cong \sqrt{\bar{n}}$; in the radio limit where hf<<kT and $\bar{n} >> 1$, $\sigma_n \cong \bar{n}$.

With these equations we may now evaluate $NEP_R(f)$, the *radiation noise-equivalent power*. Consider a black, perfectly absorbing patch of area A upon which blackbody radiation of temperature T^oK is incident within a solid angle of Ω steradians. As shown further below, the rms fluctuation in power is:

$$\sigma_p(\text{watts}) = \left(n_s \sigma_n^2\right)^{1/2} hf/\tau \tag{2.4.41}$$

where n_s is the number of states in a cavity in Bτ, B is radio frequency bandwidth (Hz), and τ is the time interval of interest, which later will be linked to the reciprocal of post-detection bandwidth.

To find σ_p we must first find n_s, the number of states impacting the sensor within Bτ.

$$n_s = (\text{electromagnetic modes in } A\Omega) \bullet \left(\frac{\text{degrees of freedom}}{\text{electromagnetic mode}}\right) \bullet \left(\frac{\text{energy states}}{\text{degrees of freedom}}\right) \tag{2.4.42}$$

The number of electromagnetic modes m in AΩ can be found by multiplying AΩ by the number of modes $ster^{-1}$ m^{-2}. The desired modal density can be found by dividing the expression (2.1.27) for intensity $I_o(f,\theta,\phi)$ by power spectral density in a single-mode TEM line: $P_+(f)$ in watts/Hz (2.1.14). This results in:

$$m = A\Omega[\text{ster m}^2]\left(2f^2/c^2\right)[\text{modes ster}^{-1}\ \text{m}^{-2}] = A\Omega 2f^2/c^2 \quad [\text{modes}] \tag{2.4.43}$$

Each of these m modes has a $2B\tau$ degrees of freedom, corresponding to 2B pulses per second times τ seconds. Finally each energy state hf has two degrees of freedom: sin ωt and cos ωt. Substituting these expressions in (2.4.42) results in:

$$n_s = m(2B\tau)(1/2) = \sin(f/c)^2\, A\Omega(2B\tau) \tag{2.4.44}$$

The variance in σ_E^2 of the cavity energy is the product of the number n_s of states at hf in $B\tau$ times the variance of state energy:

$$\sigma_E^2 = n_s \bullet \left[\sigma_n^2 \bullet (hf)^2\right] \tag{2.4.45}$$

The variance in arriving photon power averaged over time interval τ is:

$$\sigma^2 = \sigma_E^2/\tau^2 = n_s\sigma_n^2\,(hf/\tau)^2 = 2A\Omega(hf/c)^2\, B\tau\left(\bar{n} + \bar{n}^2\right)(hf/\tau)^2\,[\text{Watts}]^2 \tag{2.4.46}$$

which is the "quantum limit" imposed by intrinsic photon fluctuations associated with the radiant energy incident upon the bolometer. If the bolometer has a boxcar integrator h(t) of duration 0.5 seconds, this corresponds to a 1-Hz post-detection bandwidth; if we substitute $\tau = 0.5$, we obtain:

$$NEP_R(f) = \sqrt{4A\Omega\left(\frac{f}{c}\right)^2 B\left(\bar{n} + \bar{n}^2\right)(hf)2} \quad [\text{W Hz}^{-1/2}] \tag{2.4.47}$$

Expression (2.4.47) is the noise-equivalent power spectral density due to radiation noise.

If we are viewing a black body at all frequencies, then (2.4.17) should be integrated over all frequencies to yield $NEP_{R\infty}$, where:

$$NEP_{R\infty} = \left[4A\Omega \int_0^\infty \left(\frac{f}{c}\right)^2 (hf)^2 \left(\bar{n} + \bar{n}^2\right) df\right]^{1/2} \tag{2.4.48}$$

To perform the integral over frequency, recall:

$$\bar{n}(f) = \frac{S_I}{S_o} = \left(\frac{e^{-hf/kT}}{1 - e^{-hf/kT}}\right) \tag{2.4.49}$$

therefore, after some computation, it follows that:

$$NEP_{R\infty} = \left[4A\Omega(4kT)\frac{\sigma_{SB}}{\pi}T^4 \right]^{1/2} \quad [W\ Hz^{-1/2}] \tag{2.4.50}$$

where σ_{SB} is the *Stefan-Boltzmann constant* and equals 5.67×10^{8} $Wm^{-2}K^{-4}$, and T is the equivalent temperature of the blackbody radiation arriving at the detector. For the infinite-bandwidth case (2.4.50) can be simplified further by integrating Planck's law over frequency to find P_r, which is the total power radiated by a black body characterized by A and T and radiating at normal incidence into a small solid angle Ω.

$$P_r = A\Omega\frac{\sigma_{SB}}{\pi}T^4 \text{ Watts} \tag{2.4.51}$$

$$= A\sigma_{SB}T^4 \quad \text{if } \Omega=2\pi \tag{2.4.52}$$

Therefore, for the infinite bandwidth case and normal incidence, (2.4.50) becomes

$$NEP_{R\infty} = (P_r 16kT)^{1/2} \quad [W\ Hz^{-1/2}] \tag{2.4.53}$$

Therefore to minimize $NEP_{R\infty}$ we need to minimize P_r, and therefore detector temperature T and detector area A. As a simple example consider the radiation noise associated with a detector of area A equal $1\text{-}mm^2$ operating at liquid helium temperatures and facing 2π steradians at 4K. If we substitute $\Omega = \pi$ into (2.4.50), where a factor of 1/2 corrects for integration of (2.4.50) over 2π steradians, we obtain $NEP_{R\infty} \cong 2\times 10^{-16}$ $W\ Hz^{-1/2}$. Since for most detectors some portion of their exposed solid angle or spectral band originates from surfaces at higher temperatures, such values near 10^{-16} represent a practical lower bound; NEP's near 10^{-15} $W\ Hz^{-1/2}$ are more common, and total system NEP's near $10^{-13} - 10^{-15}$ are still more common when Johnson, phonon, and photon noises are superimposed.

2.4.6 Optical superheterodyne receivers

Optical superheterodyne receivers translate optical signals to low frequencies where they can be amplified with low-noise amplifiers, as suggested in Figure 2.4-19. We assume that S signal and P local oscillator photons/sec arrive at the photodetector, or "mixer", which has a quantum efficiency of η. For these two beams to interact they must have the same polarization at the detector. In most systems a 3-dB loss is incurred at the beam splitter because the frequencies of the two beams are usually too close together to permit a dichroic beam splitter to be used. For reasons explained later, the bandwidth B of the i.f. amplifier is usually quite narrow. The output of the mixer may be amplified and then used for communications purposes,

or squared and averaged to produce a power spectral density measurement at one or more frequencies.

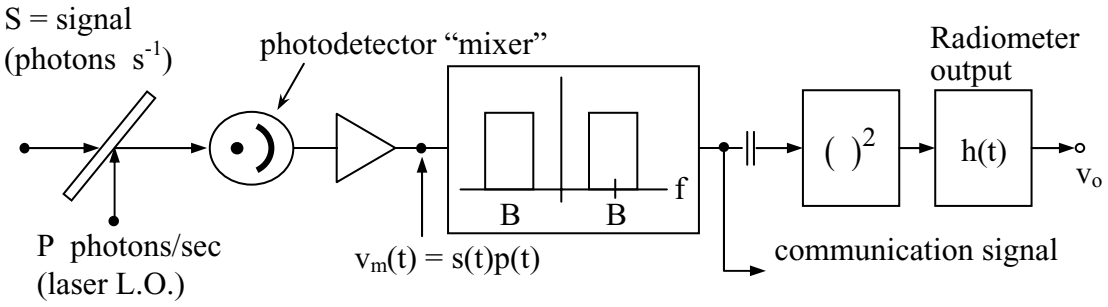

Figure 2.4-19 Optical superheterodyne circuit

In the absence of a local oscillator the mixer output in equivalent photons/sec detected is $\eta S + D$ where D is the number of photons/sec detected in the dark. This *dark count* arises from thermal excitations, cosmic rays, radioactivity, and related sources. If we represent the signal s(t) and local oscillator p(t) as monochromatic sources, as from a laser, where $s(t) = \sqrt{2S}\,\cos\omega_s t$, then $p(t) = \sqrt{2P}\,\cos\omega_o t$ and their sum squared is $\left(\sqrt{2S}\,\cos\omega_s t + \sqrt{2P}\,\cos\omega_o t\right)^2$ and:

$$v_m(t) = \text{constant}\left[\eta\left(S+P+\sqrt{SP}\,\cos\omega_{i.f.}t\right)+D\right] \tag{2.4.54}$$

where $v_m(t)$ is the number of amplified counts from the photodetector and $\omega_{i.f.} = |\omega_s - \omega_o|$. Since S is typically negligibly small compared to D, the local oscillator's strength P is chosen so that $\eta\sqrt{SP} >> D$.

If the constant in (2.4.54) is unity, then the units of $v_m(t)$ are counts/sec, which may be divided into signal and noise components:

$$v_m[\text{signal}] \cong \eta\sqrt{SP}\,\cos\omega_{i.f.}t \tag{2.4.55}$$

$$v_m[\text{rms noise}] \cong \sqrt{2\eta P}\left(\frac{\text{counts/sec}}{\sqrt{\text{Hz}}}\right) \tag{2.4.56}$$

where we have assumed the output has been smoothed by a 0.5-second boxcar integrator (e.g. by the bandpass filter B) to yield the units of $\text{Hz}^{-1/2}$, and the noise approaches $\sqrt{2D}$ if $D >> \eta P$. In general the integration time τ of the boxcar integrator h(t) at the output for Figure 2.4-19 is not

0.5 sec. If the mixer output is filtered by a bandpass filter of width B Hz and then squared to produce a normalized radiometer output v_o (see Figure 2.4-19), then:

$$v_o = \eta^2 SP\left\langle \cos^2 \omega_{i.f.} t \right\rangle \tag{2.4.57}$$

$$v_{o_{noise}} = \left(\sqrt{2\eta PB}\right)^2 \Big/ \sqrt{B\tau} \text{ for P>>S, P>>D} \tag{2.4.58}$$

where τ is the duration of the boxcar filter h(t). The carrier-to-noise ratio CNR for this optical superheterodyne radiometer is approximately the ratio of (2.4.57) and (2.4.58):

$$CNR = \eta^2 SP \Big/ \left[\eta PB / \sqrt{B\tau} \right] = \eta S\sqrt{\tau / B} \tag{2.4.59}$$

Note that the carrier-to-noise ratio in the limits where P>>D is independent of P and D, and increases with τ and decreases with B. That is, large bandwidths B pass more local oscillator noise. Equation (2.4.59) suggests we might want $\tau >> 1/B$ for further noise smoothing, but since we normally choose $\tau \cong 1/B$, (2.4.59) approaches an asymptotic optimum:

$$CNR \tilde{<} \eta S\tau \tag{2.4.60}$$

By merely increasing S and τ the CNR can be made almost arbitrarily large, independent of traditional noise sources; this is the principal advantage of optical superheterodyne receivers. However the conditions under which they are superior are limited.

If a simple detector were employed, the output associated with the signal (normalized to photon counts) would be:

$$v_{o_{sig}} = \eta S \text{ (gain normalized)} \tag{2.4.61}$$

$$v_{o_{noise\ RMS}} = \sqrt{2DB} \Big/ \sqrt{B\tau} \tag{2.4.62}$$

In this case:

$$CNR = \eta S\sqrt{\tau / 2D} \tag{2.4.63}$$

This can be compared to the superheterodyne CNR:

$$CNR_{S.H.} = \eta S\sqrt{\tau / B} \tag{2.4.64}$$

Therefore a superheterodyne yields better CNR if the i.f. bandwidth B < 2D. This can be highly constraining since the dark count in heavily cooled detectors can be quite low. On the other hand, if this dark count is quite high, the bandwidth B can be correspondingly large before the shot noise associated with the local oscillator dominates. Of course, in some systems the dark count D may be overwhelmed by other noises contributed by the mixer or i.f. amplifier.

It is interesting to compare a high performance optical superheterodyne radiometer to the sensitivity expression for a radio total-power radiometer. In the radio case:

$$\text{CNR} = T_A / \Delta T_{RMS} = T_A \sqrt{B\tau} / T_R \quad \text{for } T_R >> T_A \tag{2.4.65}$$

The corresponding expression for an optical superheterodyne is:

$$\text{Optical superheterodyne CNR} = \eta S \sqrt{\tau / B} = \eta \left(kT_A B / hf \right) \sqrt{\tau / B} \tag{2.4.66}$$

where the photons/sec S in the radio regime equals the total power received $kT_A B$ divided by the energy per photon hf. Equating these two expressions for CNR suggests the equivalent receiver noise temperature T_R for the ideal optical superheterodyne is $hf / k\eta$, which is the radio quantum limit. Thus optical superheterodyne radiometers can reach the quantum limit. Of course the received power in an optical system does not equal $kT_A B$.