22. Lecture, 16 November 1999

22.1 Coherent detection

Another way to detect light using photodetectors is to use the same method your radio uses: *coherent*, or *linear*, detection. In this method the *wave* properties of light are used explicitly, and the measurements amount to the determination of the amplitude and phase of the electric and magnetic fields in the radiation emitted by the distant source. Coherent detection itself comes in two forms that turn out to have the same sensitivity in ideal systems:

1. *coherent preamplification*, in which incident light is passed through a medium that can impart *gain*, amplifying the wave amplitudes directly. The output of such a preamplifier is usually detected by a heterodyne receiver (see below), though an incoherent detector could be used instead; the idea is for the gain to be so large that any detector could be used subsequently without affecting the signal-to-noise ratio. The paradigm for astronomical coherent preamplifiers is the *maser*, and the basic principles involved are those that apply to the (by now) more familiar *oscillator* forms of masers and lasers. We will discuss these devices superficially below. In the past decade the highest frequencies at which transistors can be used as coherent preamplifiers have crept up to tens of GHz (wavelengths down to 1 cm or so) with the development of HEMTs (see §18.8); these components are currently used in most radio-astronomical coherent preamplifiers.

2. *heterodyne detection*, in which the signal one wishes to measure is *mixed* with coherent light (constant frequency, phase and amplitude) before shining on the detector. The additional coherent light, which comprises a frequency and phase reference, is called the *local oscillator* (LO). The detector used here has to have a response time short enough that currents can exist in it at the frequency *difference* between the signal and LO: that is, at the frequency of *beats* between the signal and LO waves. Amplitudes and frequencies of the beats can be measured by the normal techniques of low-frequency electronics. In what follows we will use ν as the symbol for signal and LO frequencies (light), and f for the lower, beat, frequencies (currents).

   Heterodyne detection is used commonly – practically universally – by astronomers at wavelengths longer than about 1 mm , so we will discuss this technique in detail.

22.2 Sensitivity of heterodyne detection

Consider a photodetector used in heterodyne mode in a telescope system similar to that used in §21 (see Figure 22.1). Suppose that the beam of LO light is *matched* to the signal beam (that is, has the same beam waist size and wavefront curvature) and is injected into the signal path by use of a *diplexer* that attenuates the signal and background power negligibly. In most applications the power available from the local oscillator is many orders of magnitude larger than the signal and background power, so the diplexer can consist simply of a thin dielectric beamsplitter that transmits virtually all of the incident light but still reflects enough LO; we will assume that this is the case in the following. Note, however, that there are many instruments in which more complicated schemes such as Michelson interferometers or folded Fabry-Perot interferometers are used to combine the beams. At the surface of the photodetector (z = 0), the field from the LO is

$$\boldsymbol{E}_{LO} = \boldsymbol{E}_{0LO} e^{-i\omega_{LO} t} \quad , \tag{22.1}$$

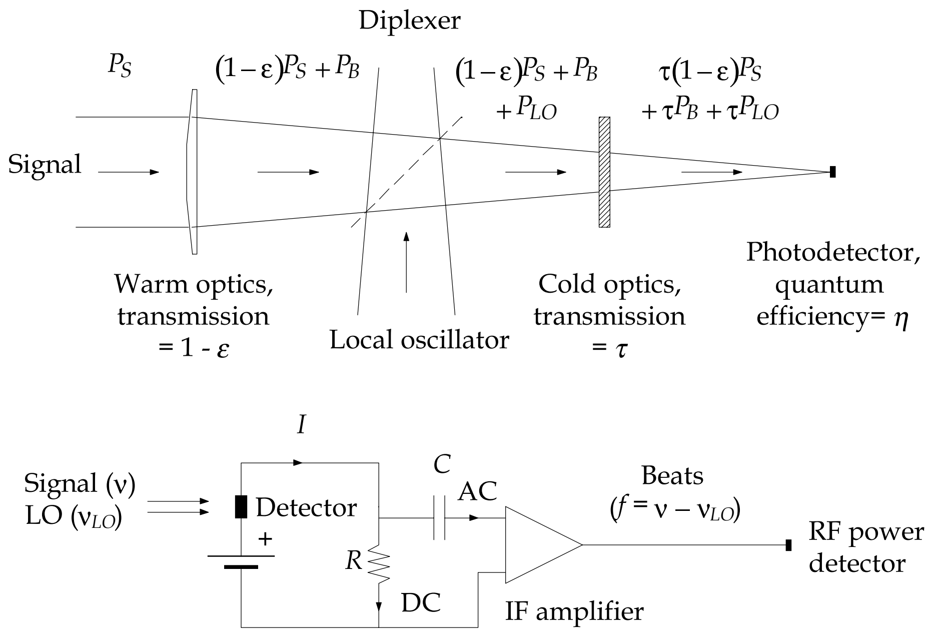

Figure 22.1: signal chain for calculation of sensitivity of a heterodyne receiver.

where $\omega_{LO} = 2\pi\nu_{LO}$ .The signal and background are not generally monochromatic or characterized by constant phase; however, we can consider for now one frequency component and phase of the signal,

$$\boldsymbol{E_S} = \boldsymbol{E}_{0S} e^{-i(\omega t + \phi)} \quad , \tag{22.2}$$

at the detector's surface. We will consider the field amplitudes to be real. (For simplicity we can leave the background power out for now; it will return in a little while.) The photocurrent induced by these two radiation fields is simply

$$\boldsymbol{I} = \frac{\eta G q P}{\hbar\omega} \quad , \tag{22.3}$$

where $P = P_S' + P_{LO}' = \tau(1-\varepsilon)P_S + \tau P_{LO}$ is given in terms of the fields at the detector's surface by

$$P = \frac{c}{8\pi}\int_A \left|\boldsymbol{E}_S + \boldsymbol{E}_{LO}\right|^2 da \quad , \tag{22.4}$$

and where A is the detector's area in the focal plane, and cgs units are used. Suppose furthermore that the detector is uniformly illuminated by both signal and LO; then

$$\begin{aligned} P &= \frac{cA}{8\pi}\left|\boldsymbol{E}_S + \boldsymbol{E}_{LO}\right|^2 \\ &= \frac{cA}{8\pi}\left[E_{0LO}^2 + E_{0S}^2 + \boldsymbol{E}_{0LO}\cdot\boldsymbol{E}_{0S} e^{i(\omega_{LO}-\omega_S)t - i\phi} + \boldsymbol{E}_{0S}\cdot\boldsymbol{E}_{0LO} e^{-i(\omega_{LO}-\omega_S)t + \phi}\right] \\ &= \frac{cA}{8\pi}\left[E_{0LO}^2 + E_{0S}^2 + 2\boldsymbol{E}_{0LO}\cdot\boldsymbol{E}_{0S}\cos\left(\left[\omega_{LO}-\omega_S\right]t - \phi\right)\right] \quad . \end{aligned} \tag{22.5}$$

Assume that the same polarization if used for signal and LO, and define

$$P_{0S} = \frac{cA}{8\pi} E_{0S}^2 \qquad P_{0LO} = \frac{cA}{8\pi} E_{0LO}^2 \quad ; \tag{22.6}$$

then we can write

$$P = P_{0LO} + P_{0S} + 2\sqrt{P_{0S} P_{0LO}} \cos([\omega_{LO} - \omega_S]t - \phi) \quad . \tag{22.7}$$

The first two terms in Equation 22.7 give rise to DC photocurrents, and the third term is the *beat* between signal and LO (see Figure 22.2), which oscillates at the *intermediate frequency* (IF), $f = \nu_{LO} - \nu_S$. Usually the detector is followed by an amplifier that works only on the IF component of this current,

$$i(\omega_S, t) = \frac{2\eta G q \sqrt{P_{0S} P_{0LO}}}{\hbar \omega_S} \cos([\omega_{LO} - \omega_S]t - \phi) \quad , \tag{22.8}$$

or an associated voltage $i(\omega_S, t)R$, as shown in Figure 22.1. The power detected at the output of the IF amplifier is proportional to the IF electrical power dissipated in the resistor *R*, or

$$i^2 R = \frac{4\eta^2 G^2 q^2 P_{0S} P_{0LO} R}{\hbar^2 \omega_S^2} \cos^2([\omega_{LO} - \omega_S]t - \phi) \quad . \tag{22.9}$$

Since it oscillates periodically, the average of this IF power over a large number of beat oscillation periods is the same as the average over a single period, $2\pi / (\omega_{LO} - \omega_S)$:

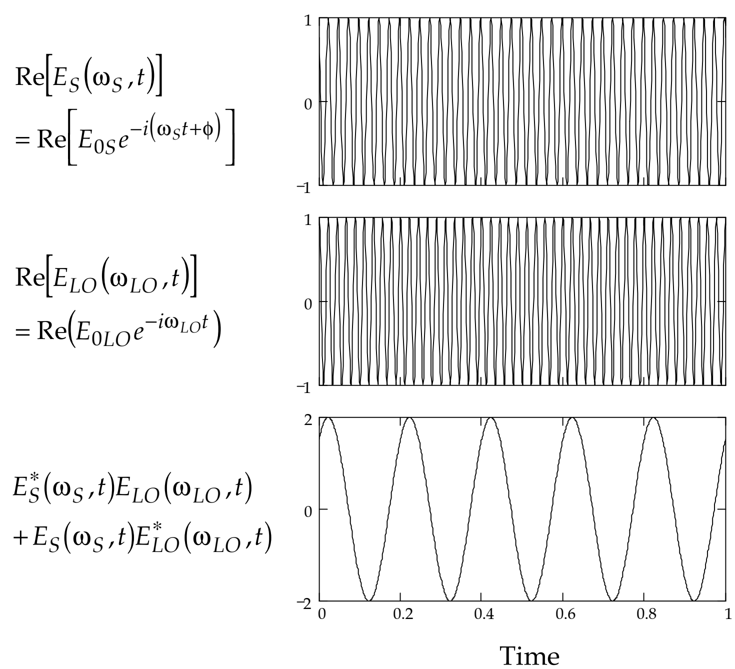

Figure 22.2: beats.

$$\begin{aligned}\overline{i^2R} &= \frac{4\eta^2G^2q^2P_{0S}P_{0LO}R}{\hbar^2\omega_S^2}\frac{\omega_{LO}-\omega_S}{2\pi}\int_0^{2\pi/(\omega_{LO}-\omega_S)}\cos^2\left(\left[\omega_{LO}-\omega_S\right]t-\phi\right)dt \\ &= \frac{2\eta^2G^2q^2P_{0S}P_{0LO}R}{h^2\nu_S^2} \quad .\end{aligned} \tag{22.10}$$

In terms of power at the system input rather than at the detector surface, this is (see Figure 22.1):

$$\overline{i^2R} = \frac{2(1-\varepsilon)\tau^2\eta^2G^2q^2P_SP_{LO}R}{h^2\nu_S^2} \quad . \tag{22.11}$$

This time, we consider our electrical signal to be a power, rather than a current.

Note that because $\cos\left[(\omega_{LO}-\omega_S)t\right]=\cos\left[(\omega_S-\omega_{LO})t\right]$, a frequency f in the photocurrent corresponds to the detection of *two* signal frequencies, $\nu_S=\nu_{LO}\pm f$, that therefore cannot be told apart simply from the IF signal. Frequencies of detected light greater than that of the LO are called the *upper sideband*, and lower frequencies are called the *lower sideband*. Separation of the two sidebands generally requires additional, interferometric optics to transmit one or the other, and this is desirable if, for instance, a spectral line is observed in one sideband, and one would like to avoid the additional noise from detection of the other sideband. In the following we will restrict our attention to heterodyne systems that detect both sidebands, and are called *double-sideband receivers.*

Now we shall deal with the noise. We assume again that the IF amplifier is designed to render Johnson noise negligible compared to shot noise, so the IF noise power, at the input of the IF amplifier, is, from Equation 21.7,

$$\left(\overline{\Delta i^2R}\right)_{\text{sn}} = I_N^2R = 2\beta GqIR\Delta f \quad , \tag{22.12}$$

where I is the average *total* current in the detector, and the label "sn" just stands for shot noise. Here follows the subtle trick of heterodyne detection: suppose that the LO power on the detector is by far the largest component of the total power:

$$I \cong I_{LO} = \frac{\tau\eta GqP_{LO}}{h\nu_{LO}} \quad , \tag{22.13}$$

so

$$\left(\overline{\Delta i^2R}\right)_{\text{sn}} = 2\beta Gq\frac{\tau\eta GqP_{LO}}{h\nu_{LO}}R\Delta f \quad . \tag{22.14}$$

Let us assume that the signal and LO frequencies are very similar $(\omega_{LO}-\omega_S \ll \omega_S)$. Then the (double sideband) signal-to-noise ratio is

$$\begin{aligned}\left(\frac{S}{N}\right)'' &= \frac{\overline{i^2R}}{\left(\overline{\Delta i^2R}\right)_{\text{sn}}} = \frac{2(1-\varepsilon)\tau^2\eta^2G^2q^2P_SP_{LO}R}{h^2\nu_S^2}\frac{h\nu_S}{2\beta\tau\eta G^2q^2P_{LO}R\Delta f} \\ &= \frac{(1-\varepsilon)\tau\eta}{\beta}\frac{P_S}{h\nu_S}\frac{1}{\Delta f} \quad ,\end{aligned} \tag{22.15}$$

independent of the LO power (!). Turn the LO power up high enough, increasing the LO photocurrent shot noise all the way, and eventually the signal-to-noise ratio doesn't depend upon LO power or this noise.

The form of Equation 22.14 is a good illustration of the workings of heterodyne detection, and turns out to be correct at the shortest wavelengths at which the technique is used by astronomers. It is, however, incomplete; it turns out that in using simply the LO shot noise we have omitted a noise process that is important at longer wavelengths, where heterodyne detection is used most often.

22.3 Background radiation and its fluctuations in heterodyne detection

We have only dealt so far with signal and LO power, since we had the $P_{0S} << P_{0LO}$ limit in mind all along. It is not much trouble to account also for detection of background power, because its beats with the LO would have exactly the same form as those of the signal – one would therefore expect to repeat the derivation of Equation 22.11, changing S for B – and because the beats between signal and background would be negligibly small compared to the beats between either with the much more powerful LO. Thus Equation 22.11 becomes (see Figure 22.1)

$$\overline{i^2 R} = \frac{2\tau^2\eta^2 G^2 q^2 P_{LO} R}{h^2 \nu_S^2}\left[(1-\varepsilon)P_S + P_B\right] \quad . \tag{22.16}$$

Just as is the case for direct detection, the form of this expression shows that we have to make *two* measurements, in practice, to determine P_S for a celestial object: one with the telescope pointing at the object, which leads to power at the input to the IF amplifier given by

$$\left(\overline{i^2 R}\right)_1 = \frac{2\tau^2\eta^2 G^2 q^2 P_{LO} R}{h^2 \nu_S^2}\left[(1-\varepsilon)P_S + P_B\right] \quad , \tag{22.17}$$

and one with the telescope pointing at blank sky:

$$\left(\overline{i^2 R}\right)_2 = \frac{2\tau^2\eta^2 G^2 q^2 P_{LO} R}{h^2 \nu_S^2} P_B \quad , \tag{22.18}$$

so that the difference between the two measured powers is proportional to the quantity we're actually trying to measure, P_S:

$$\left(\overline{i^2 R}\right)_1 - \left(\overline{i^2 R}\right)_2 = \frac{2(1-\varepsilon)\tau^2\eta^2 G^2 q^2 P_{LO} R}{h^2 \nu_S^2} P_S \quad . \tag{22.19}$$

Thus the background can be separated from the signal. However, it is essentially always the case that $P_B >> P_S$ (as well as $P_{LO} >> P_B$), since interesting astronomical objects are faint. This results in a contribution to the noise by background radiation, and this cannot be subtracted off.

That the power in blackbody radiation must fluctuate was really shown above (§20.3), when we discussed the limiting cases of the photon probability distribution; we need merely flesh out this claim here. Suppose a single-mode beam is used, and a single polarization (the LO's) is selected; then

$$P_B = \overline{P_B} = \frac{1}{2}\varepsilon B_\nu(T)\Delta\nu A\Omega = \varepsilon\frac{h\nu^3}{c^2}\overline{N}\Delta\nu\lambda^2 = \varepsilon h\nu\Delta\nu\overline{N} \quad , \tag{22.20}$$

where as usual $\overline{N} = \left(e^{h\nu/kT} - 1\right)^{-1}$. The average value of P_B^2 is, analogously,

$$\overline{P_B^2} = (\varepsilon h\nu\Delta\nu)^2\overline{N^2} \quad , \tag{22.21}$$

and the variance of the background power is

$$\overline{\Delta P_B^2} = (\varepsilon h\nu\Delta\nu)^2\overline{\Delta N^2} = (\varepsilon h\nu\Delta\nu)^2\overline{N}\left(\overline{N}+1\right) = \overline{P_B}\varepsilon h\nu\Delta\nu\left(\overline{N}+1\right) \quad , \tag{22.22}$$

where we have used Equation 20.34 in the last step. At submillimeter wavelengths and longer $(\lambda \geq 350\mu\mathrm{m})$, and common ambient temperatures (T ~ 300 K), $\overline{N}$ is considerably greater than unity, so

$$\overline{\Delta P_B^2} \cong (\varepsilon h\nu\Delta\nu)^2\overline{N}^2 = \overline{P}_B^2 \quad , \tag{22.23}$$

or

$$\left(\Delta P_B\right)_{\mathrm{rms}} = \overline{P}_B \quad ; \tag{22.24}$$

that is, the background power at these wavelengths follows Gaussian statistics, and the rms fluctuations, far from being small, are as large as the average background power itself.

22.4 The quantum limit to heterodyne detection

If our heterodyne detector can detect the background, then it can detect the background power fluctuations characterized by Equation 22.24 as well. The detected fluctuations are another form of noise, and need to be added to the shot noise power (Equation 22.13) in order to obtain a correct form for the signal-to-noise ratio. This time, however, the noise is not simply due to the finite charge on the electron; as we'll se below (§23.2), it is due to the uncertainty principle.

From Equation 22.16 we see that the background power fluctuations detected by our heterodyne receiver give rise to electrical power, referred to the input of the IF amplifier, of

$$\left(\overline{\Delta i^2 R}\right)_{\mathrm{bf}} = \frac{2\tau^2\eta^2G^2q^2P_{LO}R}{h^2\nu_S^2}\left(\Delta P_B\right)_{\mathrm{rms}} = \frac{2\tau^2\eta^2G^2q^2P_{LO}R}{h^2\nu_S^2}\overline{P}_B \quad . \tag{22.25}$$

We should add this term to Equation 22.13 to get the total IF noise power, referred to the amplifier imput:

$$\begin{aligned}\left(\overline{\Delta i^2 R}\right) &= \left(\overline{\Delta i^2 R}\right)_{\mathrm{sn}} + \left(\overline{\Delta i^2 R}\right)_{\mathrm{bf}} \\ &= \frac{2\beta\tau\eta G^2q^2P_{LO}R}{h\nu_S}\Delta f_{IF} + \frac{2\tau^2\eta^2G^2q^2P_{LO}R}{h^2\nu_S^2}\overline{P}_B \quad .\end{aligned} \tag{22.26}$$

Note that since the frequency of the LO is fixed, the signal bandwidth $\Delta\nu$ is equal to the IF bandwidth Δf_{IF} for a single sideband, or $2\Delta f_{IF}$ for double-sideband response. We will continue to assume that a double-sideband receiver is used, for which Equations 22.20 and 22.20 therefore give

$$\begin{aligned}\left(\overline{\Delta i^2 R}\right) &= \frac{2\beta\tau\eta G^2 q^2 P_{LO} R}{h\nu_S}\Delta f_{IF} + \frac{4\tau^2\eta^2 G^2 q^2 P_{LO} R}{h^2\nu_S^2}\varepsilon h\nu_S \Delta f_{IF}\overline{N} \\ &= \left(\frac{2\tau^2\eta^2 G^2 q^2 P_{LO} R}{h^2\nu_S^2}\right)\left(\frac{\beta h\nu\Delta f_{IF}}{\tau\eta}\right)\left(1+2\frac{\tau\eta}{\beta}\varepsilon\overline{N}\right) \quad .\end{aligned} \tag{22.27}$$

The IF signal power is given by Equation 22.19, and the double -sideband signal-to-noise ratio is

$$\begin{aligned}\left(\frac{S}{N}\right)' &= \frac{\overline{i^2 R}}{\overline{\Delta i^2 R}} = \frac{(1-\varepsilon)\dfrac{2\tau^2\eta^2 G^2 q^2 P_{LO} R}{h^2\nu_S^2}P_S}{\left(\dfrac{2\tau^2\eta^2 G^2 q^2 P_{LO} R}{h^2\nu_S^2}\right)\left(\dfrac{\beta h\nu\Delta f_{IF}}{\tau\eta}\right)\left(1+2\dfrac{\tau\eta}{\beta}\varepsilon\overline{N}\right)} \\ &= (1-\varepsilon)\frac{\tau\eta}{\beta}\frac{P_S}{h\nu_S}\frac{2}{\Delta\nu}\frac{1}{1+2\dfrac{\tau\eta}{\beta}\varepsilon\overline{N}} \quad .\end{aligned} \tag{22.28}$$

The only difference between this expression and the "incomplete" Equation 22.15 is the last factor. Note that since $\overline{N} >> 1$, as it is at long wavelengths and common ambient temperatures, this factor can reduce the signal-to-noise ratio significantly if $\tau\eta/\beta$ and ε are large enough. Remember that the factor of two accompanying the $\varepsilon\overline{N}$ factor is from the assumption that background fluctuations were detected in both sidebands, and that $\Delta\nu = 2\Delta f_{IF}$; this factor goes away, and $\Delta f_{IF} = \Delta\nu$, for single sieband receivers

Normally in radio astronomy the bandwidth is on the order of $\Delta f_{IF} = \Delta\nu/2 = 0.01-1$ MHz, corresponding to integration time on the order of $\Delta t' = 1/2\Delta f_{IF} = 0.5-50\ \mu$s. As you might imagine, one normally averages for much longer than a small fraction of a second. As we have seen repeatedly, fluctuations "integrate down" – become smaller – in proportion with the square root of the exposure time. Thus if the signal and noise are averaged over an exposure time $\Delta t >> \Delta t'$, the signal-to-noise ratio increases by the factor $\sqrt{\Delta t/\Delta t'}$, which gives

$$\frac{S}{N} = \left(\frac{S}{N}\right)'\sqrt{\frac{\Delta t}{\Delta t'}} = (1-\varepsilon)\frac{\tau\eta}{\beta}\frac{P_S}{h\nu_S}\frac{1}{1+2\dfrac{\tau\eta}{\beta}\varepsilon\overline{N}}\sqrt{\frac{2\Delta t}{\Delta f_{IF}}} \quad . \tag{22.29}$$

This is the complete expression for the signal-to-noise ratio of an ideal, *quantum-noise limited* double-sideband heterodyne receiver. As we did for incoherent detection, we can define a noise equivalent power for the ideal heterodyne receiver. This would be the value of $P_S/\sqrt{\Delta f}$ that corresponds to *S/N* = 1, with $\Delta f = 1/2\Delta t$:

$$1 = (1-\varepsilon)\frac{\tau\eta}{\beta}\frac{NEP}{h\nu_S}\frac{1}{1+2\dfrac{\tau\eta}{\beta}\varepsilon\overline{N}}\sqrt{\frac{1}{\Delta f_{IF}}} \quad , \tag{22.30}$$

or

$$NEP = \frac{h\nu_S}{(1-\varepsilon)\dfrac{\tau\eta}{\beta}}\left(1+2\frac{\tau\eta}{\beta}\varepsilon\overline{N}\right)\sqrt{\frac{\Delta\nu}{2}} \quad . \tag{22.31}$$

The same factor-of-two differences between double-sideband (used here) and single-sideband response that we noted in connection with Equation 22.28 also apply here.

