Submillimeter Astronomy

1. Introduction

The submillimeter regime represents a transition between infrared and radio methods. Incoherent detectors are optimum for observations at low spectral resolution. However, since no high-performance photo-detectors are known, bolometers come to center stage. For high-resolution spectroscopy (and interferometers), coherent detectors are used.

1.1 Bolometers

The photon detectors that are the basis of most of the instruments discussed so far have common signal and noise properties: 1.) the signal is derived from charge carriers freed by the absorption of a photon and the release of its energy; 2.) the noise in the signal is, to first order and in the ideal case, proportional to the square root of the number of free charge carriers; 3.) the fundamental detector noise results from processes other than photo-absorption that also free charge carriers; and 4.) they are also subject to Johnson noise and kTC noise, which result from the thermodynamics of simple electrical circuits. For bolometers, the absorbed photon energy is thermalized and heats the detector, which is connected to a heat sink through a poorly conducting thermal link (Figure 1). The resulting change of detector temperature is sensed to produce a signal. As a result, the signal is proportional to the absorbed energy, not the number of photons absorbed. The performance of the detector for small signals is characterized by the noise equivalent power, NEP, which is the power onto the detector that produces a signal into electronics with a 1 Hz bandwidth that is equal to the noise into the same electronics. The signal noise still has a component proportional to the square root of the number of absorbed photons:

\[ NEP_{ph} = \frac{hc}{\lambda} \left( \frac{2\varphi}{\eta} \right)^{1/2} \]  

(1)

where \( \varphi \) is the photons s\(^{-1} \) and \( \eta \) is the fraction of these photons that are absorbed (the quantum efficiency). The noise also includes Johnson noise (because bolometers are not easily adaptable to integrating, kTC noise is irrelevant). However, by operating them very cold, this noise can be made negligible.

However, there is also a fundamental noise associated with the storage of energy as a result of the heat capacity of the detector and the fluctuations of entropy across the thermal link that result in release of this energy at a variable rate (van der Ziel 1976, Mather 1982, Rieke 2003):
\[ \text{NEP}_r = \left( \frac{4kT^2 G}{\eta} \right)^{1/2} \quad (2) \]

where \( T \) is the temperature of operation and \( G \) is the strength of the thermal link, in watts per degree (K).

The speed of response of a bolometer is set by the ratio of the heat capacity of its detector to the strength of the thermal link:

\[ \tau_T = \frac{C}{G} \quad (3) \]

Even though the realized speed of the detector can be increased through feedback from the bias circuit (discussed below), high performance with adequate time response demands low heat capacity. The specific heats of bolometer materials - dielectrics, semiconductors, and metals - all decrease with decreasing temperature. Bolometers are operated at very low temperature both to reduce thermal noise and to reduce the response time.

A bolometer operates best when the combination of heat dissipated in the thermometer plus the heat from the infrared background raises its temperature to about 1.5 times that of its heat sink. For linear response, the power dissipated in the thermometer must exceed the infrared power. These constraints set optimum values to \( G \) and the operating temperature even when the time response is not an issue. Empirically, it is found that the achievable \( \text{NEP} \) scales approximately as \( T^{2.1 \text{ to } 2.5} \) (Rieke 2003). To achieve photon-noise-limited performance requires temperatures of \( \sim 0.3K \) on the ground and \( \sim 0.1K \) when using cold optics in space.

**1.2 Coherent Receivers**

Heterodyne receivers mix signals of different frequency; if two such signals are added together, they beat against each other. The resulting signal contains frequencies only from the original two signals, but its amplitude is modulated at the difference, or beat, frequency; this downconverted beat signal is used for the detection (see Figure 2). For heterodyne operation, the mixed field must be passed through a nonlinear circuit element or mixer that converts power from the original frequencies to the beat frequency. In the submillimeter- and millimeter-wave (and radio) region this element is a diode or other nonlinear electrical circuit component. For visible and infrared operation, the nonlinear element is a photon detector, sometimes also termed a photomixer.
If the mixer has a linear $I$-$V$ curve, then the conversion efficiency is zero (Figure 3a). Similarly, any mixer having a characteristic curve that is an odd function of voltage around the origin will have zero conversion efficiency if operated at zero bias, although a low level of efficiency can occur for operation away from zero (e.g., A in Figure 3b). Much greater efficiency is achieved, however, with a characteristic curve that is an even function of voltage (Figure 3c); if $I \propto V^2$, the output current is proportional to the square of the signal amplitude. $I \propto V^2 \propto \xi^2 \propto P$, where $\xi$ is the strength of the electric field; hence, we prefer to use square law devices as fundamental mixers because their output is linear with input power.

Assume that we are using a mixer illuminated with a mixture of two sources of power, one a signal at frequency $\omega_s$ and the other at $\omega_{LO}$. Let the power at $\omega_{LO}$ originate from the local oscillator (LO) within the instrument. We also specify that $\omega_s > \omega_{LO}$. Then the mixed signal will be amplitude modulated at the intermediate frequency $\omega_f = |\omega_s - \omega_{LO}|$. The signal at $\omega_f$ contains spectral and

Figure 2. Mixing of two signals (S1 and S2) of slightly different frequencies (a) produces a beat in the amplitude of the resulting composite signal (b)

Figure 3. I-V curves of possible mixers.
phase information about the signal at \( \omega_S \). The signal has been downconverted to a much lower frequency than \( \omega_S \) or \( \omega_{LO} \), in the frequency range where it can be processed by conventional electronics. The spectral bandwidth of a heterodyne receiver is determined by the achievable bandwidth of these intermediate frequency electronics.

Unlike incoherent detectors, this signal encodes the spectrum of the incoming signal over a range of input frequencies and also retains information about the phase of the incoming wavefront. This “extra” information allows efficient spectral multiplexing (many spectral elements observed simultaneously with a single receiver) and very flexible use of arrays of telescopes and receivers for interferometry.

Since the etendue must be preserved through the telescope, the signal photons cannot be concentrated onto the mixer in a parallel beam; even for a point source, they will strike it over a range of angles. The requirement that interference occurs at the mixer between the laser and the signal photons sets a requirement on the useful range of acceptance angle for the heterodyne receiver. This condition can be expressed as

\[
\Phi \approx \frac{\lambda}{D}, \quad (4)
\]

where \( D \) is the diameter of the telescope aperture and \( \Phi \) is the angular diameter of the field of view on the sky. Thus, a coherent receiver must operate at the diffraction limit of the telescope.

A second restriction is that the interference that produces a heterodyne signal only occurs for components of the source photon electric field vector that are parallel to the electric field vector of the laser power; i.e., only a single polarization of the source emission produces any signal. The relation between wavelength and etendue and the constraints on polarization are manifestations of the \textit{antenna theorem}, applicable to all heterodyne receivers.

There is no way of telling in the mixed signal whether \( \omega_S > \omega_{LO} \) or \( \omega_{LO} > \omega_S \). Because we have lost the initial information regarding the relative values of \( \omega_S \) and \( \omega_{LO} \), many of the derivations of receiver performance will assume that the input signal contains two components of equal strength, one above and the other below the LO frequency \( \omega_{LO} \). Since the signal at \( \omega_{IF} \) can arise from a combination of true inputs at \( \omega_{LO} + \omega_{IF} \) and \( \omega_{LO} - \omega_{IF} \), it is referred to as a double sideband signal. When observing continuum sources, the ambiguity in the

![Figure 4. LO signal and sidebands.](image)
frequency of the input signal is a minor inconvenience. When observing spectral lines, the image frequency signal at the off-line sideband results in complications.

Coherent receivers are subject to the noise due to the limited statistics in the incoming photon stream, just as with all other types of detectors. However, they have an additional source of noise, because retaining phase information is equivalent to measuring accurately the time of arrival of a photon. By the Heisenberg Uncertainty Principle, there is an unavoidable minimum noise in the measurement of both the energy and time of arrival of the photon:

$$\Delta E \Delta t \geq \frac{h}{4\pi} \quad (5)$$

From this expression, one can derive the quantum limit of a receiver:

$$T_N = \frac{h\nu}{k} \quad (6)$$

The noise temperature, $T_N$ is defined as the temperature of a black body placed over the receiver input that would be detected at signal-to-noise of 1.

It is often convenient to express the flux from a source as an antenna temperature, $T_S$, in analogy with the noise temperature. This concept is useful in the millimeter and radio regions, where the observations are virtually always at frequencies that are in the Rayleigh Jeans regime ($h\nu \ll kT$). In this case, the antenna temperature is linearly related to the input flux density:

$$\frac{P_S}{\Delta \nu} = 2kT_S \quad (7)$$

where $\Delta \nu$ is the frequency bandwidth. To maintain the simple formalism in terms of noise and antenna temperatures, it is conventional to use a Rayleigh Jeans equivalent temperature such that equation (7) holds by definition whether the Rayleigh Jeans approximation is valid or not.

The achievable signal-to-noise ratio for a coherent receiver is given in terms of antenna and system noise temperatures by the Dicke radiometer equation:

$$\left(\frac{S}{N}\right)_C = K \frac{T_S}{T_N^S} \left(\Delta f_{IF} \Delta t\right)^{1/2} \quad (8)$$

where $\Delta t$ is the integration time of the observation, $\Delta f_{IF}$ is the IF bandwidth, and $K$ is a constant of order one.

1.3 Comparison of Incoherent and Coherent Detection

Equations (7) and (8) give us the means to compare the performance of coherent and incoherent detectors, as long as we also keep in mind the bandwidth and single mode detection restrictions that we have already
discussed. From equation (6) and the definition of NEP, the signal-to-noise ratio with an incoherent detector system operating at the diffraction limit is

\[ \left( \frac{S}{N} \right)_i = \frac{2kT_S \Delta \nu (\Delta t)^{1/2}}{\text{NEP}} \]  \hspace{1cm} (9)

Therefore, using equation (8), we obtain the ratios of signal to noise achievable with the two types of system under the same measurement conditions:

\[ \frac{(S/N)_C}{(S/N)_i} = \frac{\text{NEP} (\Delta f_{IF})^{1/2}}{2kT_S \Delta \nu} \]  \hspace{1cm} (10)

Suppose a bolometer is operating background limited and we compare its signal to noise on a continuum source with a heterodyne receiver operating at the quantum limit. We set the bolometer field of view at the diffraction limit, \( A\Omega = \lambda^2 \) and assume that the background is in the Rayleigh Jeans regime (e.g., thermal background at 270K observed near 1mm). The background limited NEP is:

\[ \text{NEP} = \frac{\hbar c}{\lambda} \left( \frac{2\varphi}{\eta} \right)^{1/2} \]  \hspace{1cm} (11)

The photon incidence rate, \( \varphi \), can be shown to be

\[ \varphi = \frac{2\eta kT_B \Delta \nu}{h \nu} \]  \hspace{1cm} (12)

If we assume the bolometer is operated at 25% spectral bandwidth, \( \Delta \nu = 0.25\nu \), then

\[ \frac{(S/N)_C}{(S/N)_i} \approx \frac{2.4 \times 10^6}{\nu} (4\Delta f_{IF})^{1/2} \approx \frac{2.6 \times 10^{11}}{\nu} \]  \hspace{1cm} (13)

We have taken the IF bandwidth to be \( 3 \times 10^9 \) Hz, a typical value. Thus, the incoherent detector becomes more sensitive near \( 2.6 \times 10^{11} \) Hz and at higher frequencies, or at wavelengths shorter than about 1mm. Actually, this comparison is slightly unfair to it (since, for example, it does not have to work at the diffraction limit), so it is the detector of choice for continuum detections to wavelengths of 2 to 3mm. Hence, the development of large scale bolometer cameras for mm-wave and submm telescopes. Conversely, at wavelengths longer than a couple of millimeters, equation (13) shows why coherent detectors are the universal choice. Of course, coherent detectors are preferred for high resolution spectroscopy and for interferometry (as we will discuss later).

2. Bolometers

Very high performance bolometers have been built into small arrays for some time, but until recently
these devices were based on parallel operation of single pixels. The obstacle to true array-type construction was that the very small signals required use of junction field effect transistor (JFET) amplifiers that needed to operate above about 50K, far above the operating temperature of 0.3K or below for the bolometers themselves. It is difficult to implement the simple integration of detector and amplifier that is the heart of array construction with this temperature difference. With the development of adequately low-noise readouts that can operate near the bolometer temperature, the first true high-performance bolometer arrays for the far infrared and submillimeter spectral ranges are just becoming available. For example, one channel of the Herschel/PACS instrument uses a 2048 pixel array of bolometers (Billot et al. 2006). The architecture of this array is vaguely similar to the direct hybrid arrays for the near- and mid-infrared. One silicon wafer is patterned with bolometers, each in the form of a silicon mesh, as shown in Figure 5. The delicate construction of the detector depends on the ability to etch exquisitely complex miniature structures in silicon.

The development of "silicon micromachining" has enabled substantial advances in bolometer construction generally and is central to making large-scale arrays. In this instance, the silicon mechanical structure around the mesh region provides the heat sink; the mesh is isolated from it with thin and long silicon rods. The rods and mesh both need to be designed carefully to achieve appropriate response and time constant characteristics. The mesh is blackened with a thin layer of titanium nitride with sheet resistance matched to the impedance of free space (377Ω/square section of film). This matching provides an efficiency of 50% over a broad band in absorbing submm or mm-wave photons. Quarter-

Figure 5. A single pixel in the Herschel/PACS bolometer array, pixel size about 750μm.
wave resonant structures can tune the absorption to higher values over limited spectral bands. For each bolometer a silicon-based thermometer doped by ion implantation to have appropriate temperature-sensitive resistance lies at the center of the mesh. Large resistance values are used so the fundamental noise is large enough to utilize MOSFET readout amplifiers. A second silicon wafer is used to fabricate the MOSFET-based readouts, and the two are joined by indium bump bonding. When far infrared photons impinge on the array, they are absorbed by the grids and raise the temperatures of the thermometers. The resulting resistance changes are sensed by the readouts, amplified, and conveyed to the external electronics. To minimize thermal noise and optimize the material properties, the bolometer array is operated at 0.3K. Further details are in Billot et al. (2006).

Another approach is taken in transition edge sensor (TES) arrays such as the ones to be used in the submillimeter camera SCUBA-2. The name of these devices is derived from their thermometers, which are based on thin superconducting films held within their transition region, where they change from the superconducting to the normal state over a temperature range of a few mK. In this region, the films have a stable but very steep dependence of resistance on temperature. The temperature of this transition can be set to a convenient point by using a "bilayer" film consisting of a layer of normal material and one of superconductor. The Cooper pairs from the superconductor can diffuse into the normal metal and make it weakly superconducting, a process called the proximity effect. As a result, the transition temperature is lowered relative to that for the pure superconducting film. Adjusting the film thicknesses changes the extent of the effect and adjusts the transition temperature to a convenient value (e.g., one appropriate for a specific type of low-temperature refrigerator).

However, the resistance of a TES is low, so it can deliver significant power only to low input impedance amplifiers, which rules out JFETs and MOSFETs. Instead, the signals are fed into superconducting quantum interference devices (SQUIDs). A SQUID (Figure 6) consists of an input coil that is inductively coupled to a superconducting current loop. Two Josephson junctions - junctions of superconductors with an intervening insulator - interrupt the loop. The Cooper pair current across a Josephson junction is a sinusoidal function of the superconducting phase difference between the two sides of the junction. The superconducting phase around the current loop is also a function of the magnetic flux through the loop, and thus of the electrical current through the input coil. In a phenomenon analogous to a two-slit optical interferometer, interference of the superconducting wavefunction around the loop results in a voltage response on the output of the SQUID that is a very sensitive function of the current applied to the input coil. Thus, changes in the bolometer current produce a large modulation of the SQUID current - i.e., when its output is made linear by using feedback, the device works as an amplifier. SQUIDs are the basis for a growing family of electronic devices that operate by superconductivity.
Because of the steep temperature dependence of their resistance, TESs are most stable when biased with a constant voltage. The SCUBA-2 devices, as with all TES bolometers, are operated in this mode (see Figure 6 for a typical bias circuit). In this state, when their temperature rises due to power from absorbed photons, their resistance rises, the bias current drops, and the electrical power dissipation in them decreases, partially canceling the effects of the absorbed power and limiting the net thermal excursion. This behavior is called electrothermal feedback. The steep temperature dependence of the resistance of a TES makes the effect very strong. This feedback expedites operating arrays with TESs because minor variations in the transition temperature can be overcome by the tendency of the feedback to force each device to a suitable operating point. Electrothermal feedback can also make the bolometers operate tens or even hundreds of times faster than implied by Equation (3). In fact, if the TES is too fast, the bolometer/SQUID circuit can be unstable and measures must be taken to slow the response.

TES bolometer arrays use SQUIDs for the same readout functions that we have discussed for photodiode and IBC detector arrays. The operation of a simple SQUID time-domain multiplexer is illustrated in Figure 6. The biases across the SQUIDs are controlled by the address lines. Each SQUID can be switched from an operational state to a superconducting one if it is biased to carry about 100μA. The address lines are set so all the SQUIDs in series are superconducting except one, and then only that one contributes to the output voltage. By a suitable series of bias settings, each SQUID amplifier can be read out in turn.

From our perspective, a major advance in these devices is that the superconducting readouts operate with very low power dissipation and at the ultra-low temperature required for the bolometers.
Therefore, integration of detectors and readouts is simplified and the architecture can potentially be scaled to very large arrays.

An example is the bolometer array for SCUBA-2. Each SCUBA-2 array is made of four sub-arrays, each with 1280 transition-edge sensors. The design is illustrated in Figure 7. The detector elements are separated from their heat sinks by a deep etched trench that is bridged by only a thin silicon nitride membrane. The absorbing surface is blackened by implanting it with phosphorus to match the impedance of free space. The dimensions of the array pixels are adjusted to form a resonant cavity at the wavelength of operation, to enhance the absorption efficiency. The superconducting electronics that read out the bolometers are fabricated on separate wafers. The two components are assembled into an array using indium bump bonding.

In addition to the arrays for SCUBA-2, various forms of TES-based bolometers with SQUID readouts are under active development in a number of laboratories. For example, in the mm-wave range the pixel-based array geometry in Figures 6 and 7 can be replaced with tiny antennae defined by photolithography. The operating spectral range of antenna-coupled bolometers lies between the infrared and radio and they make use of a mixture of technologies from both spectral regimes. The antenna feeds respond to a single polarization, an advantage if the detectors are planned for a polarimeter. Antennae can be arranged in a single focal plane to measure several polarization angles simultaneously. Microstrip transmission lines can bring the antenna signals outside the sensitive area of the array (a microstrip consists of a miniature circuit trace on an insulator and over a ground plane that can be designed to have some of the characteristics of a waveguide). There, the signals can be sent to a bank of microstrip filters that separate them into multiple spectral bands. Microstrip transmission lines carrying the signals are then terminated with normally conducting metal resistors and TESs sense the temperatures of the resistors as a measure of the power received by the antennae in each band.
There are two basic approaches to multiplexing TES signals. We have described the time-domain approach, but multiplexing in the frequency domain is also possible. In this case, each TES is biased with a sinusoidally varying voltage and the signals from a number of TESs are encoded in amplitude-modulated carrier signals by summing them. They are read out by a single SQUID and then brought to room-temperature electronics that recovers each of the signals by synchronous detection.

3. Submillimeter Coherent Receivers

3.1 Basic elements

Figure 8 shows the operation of a submillimeter heterodyne receiver. The signal photons (from the telescope) are combined at the beam splitter (sometimes called a diplexer) with the local oscillator signal, and then conveyed to a mixer. The mixer downconverts the combined signal to the intermediate frequency. This signal is then amplified and sent to a detector stage where it is converted to a slowly varying dc signal. We describe the individual components of this receiver below.

In the submm and lower frequencies, tunable LOs are available. They typically start with a lower frequency oscillator and put its output through a highly non-linear circuit element (e.g., a diode). The resulting waveform has substantial power in frequency overtones, which can be isolated and made the input to an amplifier, then taken to another non-linear device, from which the frequency overtones can again be isolated and amplified. The result is that the original oscillator frequency is multiplied up to the operating frequency of the receiver. As we will see, LO power is a critical asset for a receiver, and the power that can be delivered through a multiplier chain is limited.

3.2 Mixers
In the submm and at lower frequencies the mixer is a specialized electrical element onto which the energy is concentrated by antennae, waveguides, and other non-focussing optics. In the millimeter-wave regime, the highest-performance mixers are based on SIS devices. SIS stands for "superconductor-insulator-superconductor. A sandwich of these materials produces the "diode or other nonlinear device" described above as the heart of a mixer. Its operation is illustrated in Figure 9.

SIS mixers can be used up to frequencies approaching $10^{12}$ Hz. A fundamental limit results from the ability of high frequency photons to break the Cooper pairs. For example, using lead as the superconductor imposes an upper frequency limit of about $6.5 \times 10^{11}$ Hz. The frequency range can be doubled by operating the device over both positive and negative polarity; in that case, the junction must be placed in a magnetic field to suppress Josephson tunneling near zero voltage. However, in general heterodyne receivers for very high frequencies use other types of device for their mixers, such as hot electron bolometers.

Figure 9. Operation of a SIS junction. The two superconductors (S) are shown in a pseudo-bandgap diagram; the band gap is a few meV. They are arranged on either side of a thin insulating layer (I). Panel (a) shows the device without a bias voltage. A bias has been applied in (b). When it becomes big enough to align the "valence" band on the left with the "conduction" one on the right, suddenly Cooper pairs tunnel through the insulator efficiently, causing a current. The sudden onset of this current results in a sharp inflection in the I-V curve for the device. See the plot in (c) for the overall behavior. Because the inflection in the bias curve is so sharp at $2\Delta/q$, relatively little local oscillator power is needed to get a good IF signal.

Figure 10 is a SIS mixer block. The local oscillator and the input signal are coupled into the mixers by the twin slot antenna. The IF is taken out to the right, while the volume to the left helps tune the response for efficiency (but at a specific frequency). However, the tuning required for good signal coupling generally restricts the operation to a relatively narrow spectral range (10-20%).

Using the photodetector case as an example, the mixer need not respond fast enough to track the frequency of the two input signals, so those signals just produce constant photocurrents. In addition, the photocurrent
contains a component oscillating at the intermediate frequency, \( \omega_{IF} = |\omega_S - \omega_L| \). The IF current is the heterodyne signal and has a mean-square-amplitude proportional to the product of the signal and LO power. Because the signal strength depends on the LO power, many forms of noise can be overcome by increasing the output of the local oscillator.

The conversion gain is defined as the IF output power that can be delivered by the mixer to the next stage of electronics divided by the input signal power. The ability to provide an increase in power while downconverting the input signal frequency is characteristic of quantum mixers. Classical mixers do the downconversion without gain - but the use of very low noise electronics in the GHz range of the IF signal makes it useful to carry out this operation even without gain. In fact, to achieve better stability, quantum mixers are usually operated with gains less than one also.

3.3 Amplifiers

The heterodyne signal is a low level, high frequency AC current; it needs to be amplified and converted into a slowly varying voltage that is proportional to the time-averaged input signal power. The first step in this

Figure 10. SIS junction mixer block.
process is amplification. We want to preserve as high an IF frequency range as possible, since the spectral band over which the receiver works is just 2 $\Delta f_{IF}$ (assuming we use it double sideband).

The best performance for the IF amplifier is obtained with high electron mobility transistors (HEMTs) built on GaAs, as in Figure 11. The HEMT is based on the metal-semiconductor field effect transistor (MESFET). It consists of a substrate of GaAs with an n-doped layer grown on it to form the channel, with contacts for the source and drain and a gate formed as a Schottky diode between them on this layer (a Schottky diode results from the asymmetric impedance characteristics of a junction between a metal and a semiconductor). The electron flow between source and drain in this channel is regulated by the reverse bias on the gate; as with the JFET, with an adequately large reverse bias the depletion region grows to the semi-insulating layer and pinches off the current. Because this structure is very simple, MESFETs can be made extremely small, which reduces the electron transit time between the source and drain and increases the response speed.

In the HEMT, the MESFET performance is further improved by using a heterojunction (junction between two different semiconductors with different bandgaps; see Figure 12) so the electrons flow in undoped GaAs.

To achieve this result, the MESFET is grown on heavily doped GaAlAs, whose Fermi level lies above the bottom of the conduction band in the undoped GaAs layer. Thus, the conduction electrons collect in the GaAs and flow through it under the influence of the MESFET fields. The very high mobility in undoped GaAs makes for very fast response, to $\sim 10^{11}$ Hz.

### 3.4 Detector stage
The conversion to a slowly varying output can be done by a detector stage that rectifies the signal and sends it through a low-pass filter. We would like the circuit to act as a square law detector because $<I_F^2>$ is proportional to $I_S$, which in turn is proportional to the power in the incoming signal.

To demonstrate how the detector stage achieves this goal, we solve the diode equation for voltage and expand in $I/I_0$:

$$V = \frac{kT}{q} \ln \left(1 + \frac{I}{I_0}\right) \approx \frac{kT}{q} \left[ \frac{I}{I_0} - \frac{1}{2} \left(\frac{I}{I_0}\right)^2 + \frac{1}{3} \left(\frac{I}{I_0}\right)^3 - \frac{1}{4} \left(\frac{I}{I_0}\right)^4 + \ldots \right].$$  \hspace{1cm} (14)

The first and third terms in the expansion will have zero or small conversion efficiency, and the $4^{th}$ and higher terms will be small if $I \ll I_0$. Thus, the detector stage does act as a square law device.

3.5 Spectrometers

Sometimes it is desirable to carry out a variety of operations with the IF signal itself before smoothing it. For example, imagine that the down-converted IF signal is sent to a bank of narrow bandpass fixed-width electronic filters that divide the IF band into small frequency intervals. Each of these intervals maps back to a unique difference from the LO frequency, i.e., to a unique input frequency to the
receiver. A typical filter bank may have 256 or 512 channels. A detector stage can then be put at the output of each filter, so the outputs are proportional to the power at a sequence of input frequencies, that is they provide a spectrum. In this manner the total IF bandwidth (perhaps ~ 4 GHz) is divided into a spectrum, even though only a single observation with a single receiver has been made; the process is called spectral multiplexing.

Although conceptually simple, a high performance filter bank can be an engineering challenge. The filters need to have closely matched properties and be robust against drift of those properties due to effects like temperature changes. A filter bank is also inflexible in use; the resolution must be set during design and construction. Finally, these devices are complex electronically and expensive to build.

---

Figure 15. A filter bank spectrometer. The input IF signal is divided among the bandpass filters and the output of each is processed by a detector/integrator stage. The outputs of these stages are switched sequentially to the output computer where the spectrum can be displayed.

Figure 16. An Acousto Optical Spectrometer (AOS).
if many channels are required.

An acousto optical spectrometer provides a many-channel spectrometer without the electronic complexity of a filter bank, since it divides the IF signal into frequency components without a dedicated unit for each component. In this device (see Figure 16), a piezoelectric transducer is attached to a Bragg cell, a transparent volume containing either a crystal like lithium niobate or water. When the IF signal is fed into this transducer, it vibrates to produce ultrasonic waves that propagate through the Bragg cell and produce periodic density variations. As a result, the index of refraction in the cell also varies periodically, making it act like a volume phase diffraction grating. When a near-infrared laser diode output passes through the cell, it is deflected accordingly – the zero order path is absorbed and the first order path is focused onto a CCD by camera optics. The light intensity is proportional to the IF power injected into the Bragg cell, while the deflection angle and hence position on the CCD is determined by the ultrasound wavelength. The output signal is basically the Fourier transform of the IF signal.

AOS spectrometers are capable of resolving the IF signal into more than 2000 spectral channels. To achieve this level of performance, care must be taken to design around mechanical and temperature drifts so they remain stable.

The third method to obtain a spectrum from the IF signal is to compute its autocorrelation:

$$R(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} U(t) U(t + \tau) dt$$  \hspace{1cm} (15)$$

The Fourier transform then gives the spectrum. Autocorrelators are flexible in operating parameters and very stable, since they work digitally. Their biggest disadvantage is that they can only operate over a limited bandwidth, currently of order 2 GHz.

To measure the autocorrelation, the IF signal is first digitized. High speed performance is required; the Nyquist theorem says that the digitization rate must be at least twice the IF bandwidth. Therefore, autocorrelators often digitize to only a small number of bits. As shown in Table 1, the loss of information is surprisingly modest if the gains are set optimally. Many systems only digitize to two bits, and some only to a single bit. The digitized signal is taken to an electronics circuit that imposes the necessary delays in the signal using shift registers and then combines the results to provide the autocorrelation as an output.

Because it is not possible actually to carry out the limit to infinite time in equation (15), autocorrelators produce spectra with “ringing” due to any sharp spectral features. Similar behavior was discussed with regard to Fourier transform spectrometers (Chapter 5). As in that case, these artifacts can be reduced by filtering the signal (e.g., with a “Hanning filter”), but with a loss in spectral resolution.

| Table 1. Information retained as a function of digital bits |
|-------------|-----------------|
| bits       | information     |
| 1          | 64%             |
| 2          | 81%             |
| 3          | 88%             |
| infinite   | 100%            |
4. Submillimeter Observatories

Observations from the ground in the submillimeter are extremely sensitive to the amount of water vapor overlying the observatory site. Figure 17 demonstrates that the windows at 350 and 450μm close and the one at 850μm is impaired at total water vapor levels of 2mm and above. Even observatory sites considered for other purposes to be high and dry only allow submm observations under the most favorable circumstances. For example, Figure 18 shows the distribution of water vapor over Paranal, the site of the VLT, and at an altitude of 2635m. The observations are for the zenith – even there, for half of the nights the level is above 2mm, and it reaches 0.5mm only on very rare winter nights. Therefore, the selection of sites with extremely low water vapor is critical to success for a submm observatory, requiring that the observatory be either placed at high latitude (the Antarctic has been the site of a number of successful observations) or extremely high (e.g., 5000 meters elevation for ALMA).

Wind-blown fluctuations in the water vapor content of the air over a submm telescope produce a fluctuating level of thermal emission into the beam, resulting in noise. This noise rises rapidly with decreasing frequency. For modern broad-band continuum cameras (e.g., bolometric detectors), it dominates the detector noise below about 0.5 Hz. The effect largely arises close to the telescope, due to the small scale height for water vapor. The noise can be reduced by nutating the secondary mirror at a frequency of 1 Hz or higher, so the measurements are made between source and a blank region next to the source rapidly. For small nutation angles, the two positions sample nearly the identical path through the low-lying atmosphere while modulating the source signal completely. Fitting the common-mode fluctuations and spatial behavior over the pixels of an imaging array can also be effective in reducing the impact of the sky signal.
Figure 17. Atmospheric transmission as a function of precipitable water vapor, from 0.5mm (upper left) to 1mm (upper right), 2mm (lower left) and 4mm (lower right). The 350 and 450\,\mu m (windows) are between 600 and 900 GHz, the 850\,\mu m window is centered at 350 GHz. From http://www.submm.caltech.edu/cso/weather/atplot.shtml
Figure 18. Distribution of precipitable water vapor readings from the VLT site (elevation 2635m). The histograms give the differential distributions, and the smooth curves are integral ones for which the y axis scale is appropriate. From http://www.eso.org/public/outreach/products/publ/reports/whitebook/wb20.html
Figure 19 shows the design of Bolocam, and Figure 20 shows SCUBA-2 and Figure 21 the camera portion of this instrument. Bolocam includes a helium-3 refrigerator, which along with the helium 4 dewar,

dominates the view to the left. The bolometer array is shown to the right. It is an older-generation device that uses semiconductor thermometers that are read out by JFETs. The energy is brought onto the bolometers with optimized feedhorns. The feedhorns cannot produce a filled array; bolometer arrays of this design have a grid of fields of view separated by dead zones about one field of view across. As a result, a variety of complex dithering strategies are used to produce fully sampled maps.

SCUBA-2 is about to go into service (it has had long delays largely due to manufacturing issues with its bolometer arrays). It has a huge cryogenic system (Figure 20). Inside, for shielding, the instrument is surrounded by a light-tight box held at 1K. Once inside, the beam is divided into short and long wavelength arms using a dichroic filter, so both arms view the same field on the sky. Each arm has a detector array optimized for its wavelength range (recall that the cavities around the detectors are tuned to enhance the absorption).
Figure 20. SCUBA-2. Bolometer array cameras now come in large packages.
Figure 21. The 1K box for Scuba-2. The beam enters from the telescope through the cold stop. The dichroics split the beam to allow simultaneous imaging at 450 and 850μm. This entire assembly is cooled to 1K.