The Role of Coherent Detection

J. Zmuidzinas[1]

[1]Division of Physics, Mathematics, and Astronomy
California Institute of Technology, 320–47
Pasadena, CA 91125
jonas@submm.caltech.edu

Abstract

Many interesting astronomical objects, such as galaxies, molecular clouds, PDRs, star–forming regions, protostars, evolved stars, planets, and comets, have rich submillimeter spectra. In order to avoid line blending, and to be able to resolve the line shape, it is often necessary to measure these spectra at high resolution. This paper discusses the relative advantages and limitations of coherent and direct detection for high resolution spectroscopy in the submillimeter and far–infrared. In principle, direct detection has a fundamental sensitivity advantage. In practice, it is difficult to realize this advantage given the sensitivities of existing detectors and reasonable constraints on the instrument volume. Thus, coherent detection can be expected to play an important role in submillimeter and far–infrared astrophysics well into the future.

Introduction

Coherent detection is used primarily at long wavelengths, from the radio into the far-infrared. In comparison to direct detection, coherent detection offers several important advantages, including the ability to obtain very high spectral resolution. Indeed, coherent detection will play a key role in upcoming major submm/far-IR projects and missions, such as ALMA, SOFIA, and Herschel. However, coherent detection has one fundamental disadvantage, which is a limit to sensitivity that is imposed by quantum mechanics. While this "quantum limit" does not play a significant role for (warm) ground–based or airborne telescopes, it would become an important issue for cold telescopes in space, such as the large–aperture SAFIR mission envisioned for NASA. Will there be a role for coherent detection in future submillimeter space missions beyond Herschel ?

This paper examines the case for coherent detection in the submillimeter/far–IR, following the approach of an earlier paper. [1]. The case is straightforward: first, high spectral resolution is scientifically important; and second, although in principle direct detection does have a sensitivity advantage, it is difficult to realize this advantage in practice for the case of high resolution spectroscopy.

1. Submillimeter and Far–Infrared Spectroscopy

Spectroscopy in the submillimeter and far–IR, although still in early stages of development, has the potential to become a very powerful tool for astrophysics. Here one finds the ground–state transitions of numerous hydride species that are critical for understanding interstellar chemistry. One also has numerous diagnostics of warm gas, such as the high–J CO lines as well as the atomic and ionic fine structure lines of the lighter elements such as C, N, and O. At these long wavelengths, observations can penetrate through very large column densitites of dust, which are totally opaque in the infrared, optical, and UV, and only start becoming transparent again to X–ray photons.

So far, we have collectively obtained only a small glimpse of what submm/far–IR spectroscopy has to offer, through the pioneering work done on the Kuiper Airborne observatory and ground–based telescopes, followed by the substantial advances enabled by ISO. For example, the ISO far–IR spectra of nearby galaxies [2] show a wide variety of characteristics, and issue a challenge to all of us to decode their meaning. SOFIA and the Herschel Space Observatory will make enormous contributions, pushing beyond ISO to longer wavelengths, better sensitivity, better angular resolution, and higher spectral resolution.

In the future, it will be possible go beyond Herschel by many orders of magnitude in sensitivity, by using a colder telescope, optimized instrumentation, and better detectors. Although much attention has been focused on the *detection* of distant objects through submm/far–IR imaging, sensitive submm/far-IR *spectroscopic* observations of distant, high–redshift "submillimeter" galaxies will be a crucial next step toward understanding their redshift distribution, energy sources, and origin. Such distant–object studies will probably best be done with moderate resolution direct–detection spectrometers. However, developing an understanding of the basic physical phenomena involved will require the study of nearby objects at higher spectral resolution.

2. The Importance of High Spectral Resolution

There are many examples that can be given to illustrate the importance of high spectral resolution. Submillimeter line surveys, such as the recent 650 GHz CSO survey [3], have shown that the spectra of star–forming regions can be exceedingly rich, and that high spectral resolution (1 km s^{-1} or better) is necessary to deal with line confusion and line blending. The detection and subsequent analysis of the abundance, excitation, and physical origin of numerous chemical species would be impossible without high spectral resolution.

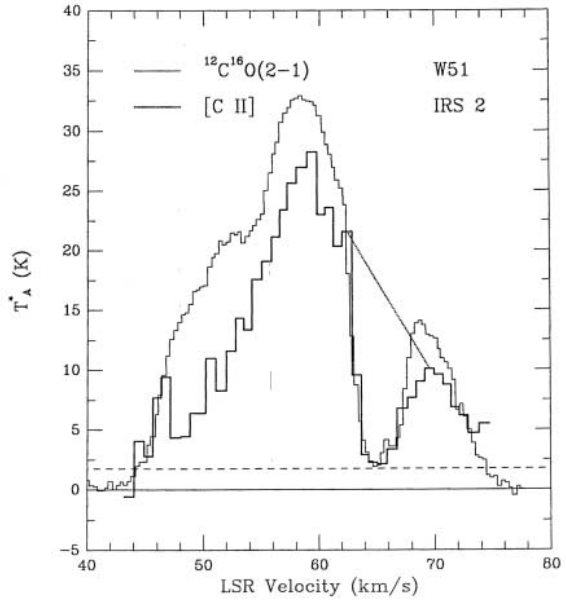

Figure 1: A high–resolution spectrum of the 158 μm [C II] line toward W51 IRS2 [4]. Note the sharp absorption feature, which in this case is most likely produced by a foreground cloud with a low–density PDR.

Water line observations from SWAS at 557 GHz provide another example. Here the line profiles are often observed to be "self–reversed", with deep, sharp absorption dips superposed on a broader emission profile [5]. This type of profile is expected to be common for the case of embedded sources, where the warmer gas located closer to the source is surrounded by cooler foreground material. The abundance of water vapor can vary greatly with location, and the large molecular dipole moment provides a strong coupling to the submm/far–

IR radiation field. These effects, along with large optical depths, can also be expected to yield complex line profiles. In such situations, comparison of optically thick and optically thin line profiles (e.g. $H_2{}^{16}O$ vs. $H_2{}^{18}O$) can lead to important insights.

Furthermore, complex profiles can occur even for species that are spatially widespread and have transition strengths and optical depths that are normally thought to be low. The 158 μm line of ionized carbon (C II) provides a good example. High resolution spectroscopy of galactic star–forming regions (performed on the Kuiper Airborne Observatory in the 1980's by Betz, Boreiko, and Zmuidzinas) often reveals complex line profiles, with deep, optically thick absorption dips (see Figure 1) due to foreground material. Ground–state transitions are particulary susceptible to this effect.

Finally, line profiles can often carry important information about the nature of the emission. A recent example is the submillimeter HCN laser in IRC+10216 [6], whose narrow line width is a characteristic signature of laser action.

3. The Connection Between Spatial and Spectral Resolution

Both spatial and spectral resolution are important. Although one often sees these plotted on orthogonal axes when describing instrument capabilities, in fact often there is a close connection:

- High spatial resolution often results in narrow line widths
- Spectral resolution can be used to gain spatial information
- Spectral resolution can help break through the broadband photometric spatial confusion limit

One example of the first point is the mapping of line emission from nearby galaxies with millimeter interferometers, where the synthesized beam covers a single molecular cloud, as compared to single–dish observations, where the beam picks up the entire rotation curve of the galaxy.

This behavior is not universal, however; in some cases the line profile becomes broader at higher spatial resolution, as occurs for galactic nuclei, or more generally when a localized gravitational potential plays an important

role in the dynamics. For some of these cases, one can use spectral information to gain spatial information. A good example is the analysis of the CO emission from the nuclear region of Arp 220 by Scoville and collaborators [7], [8], in which a self–consistent model of the mass distribution and dynamics was used to infer the structure at a scale smaller than the synthesized beam.

One does not necessarily need to have a dynamical model to extract spatial information from spectral data; a purely empirical approach is possible. In this case, one uses a second data set which has both high spatial and spectral resolution to perform the decoding from the spectral to spatial domain. There are well–defined algorithms for doing this. A recent example of this approach is the use of an IRAM CO(2-1) interferometer map along with a CSO CO(6-5) map to determine a line ratio map at somewhat higher spatial resolution than the CSO data [9]. In this context, it is intriguing to consider the use of ALMA maps in combination with high spectral resolution observations with SAFIR. For instance, what might this combination provide for the study of protoplanetary disks ?

4. Coherent Detection

Coherent and incoherent detection are easily distinguished. The fundamental difference is that coherent detection instruments respond to the complex amplitude of the field (amplitude and phase), whereas incoherent detection instruments respond to the intensity (power). Using the language of quantum mechanics, coherent instruments measure the value of a, the photon destruction operator, whereas incoherent instruments measure $a^\dagger a$, the photon number operator. The "quantum noise" limitation for coherent receivers arises from the fact that they measure both amplitude and phase (or "position" and "momentum") simultaneously, and these are not commuting obervables, and so are subject to an uncertainty principle [10]. The value of this quantum noise can be expressed as a noise temperature, $T_n = h\nu/k_B$, which is 0.05 K/GHz, or 50 K/THz. Equivalently, it corresponds to the photon shot noise from a background of 1 photon per second per Hertz of bandwidth.

In essence, coherent instruments amplify the field into the classical domain. The relevant quantity describing this translation from the quantum to the classical regime is the *photon number gain*, which is not necessarily synonymous with power gain. Figure 2 shows a block diagram of a typical heterodyne re-

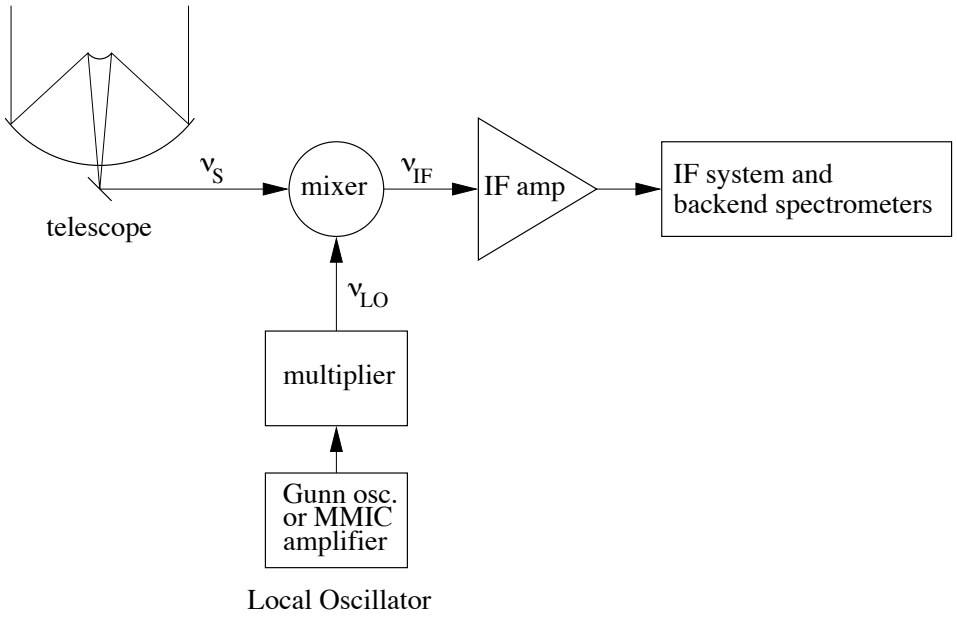

Figure 2: A block diagram of a submillimeter heterodyne receiver system. The signal from a telesope at a frequency ν_S is combined with a local oscillator at frequency ν_{LO} in a "mixer", which is a nonlinear device, to yield the difference or "intermediate" frequency $\nu_{IF} = |\nu_S - \nu_{LO}|$, which is then amplified and and spectrally analyzed.

ceiver. Even though the mixer may have power conversion loss, it most cases it has photon number gain, because the output frequency is much lower than the input frequency. Of course, the receiver system should always have overall photon number gain after the signal passes through the IF amplifier.

The physical origin of quantum noise can be illustrated for an ideal hypothetical amplifier, as shown in Figure 3. This device, consisting of an inverted population of atoms or molecules in a tube, clearly has photon number gain due to stimulated emission. It also has noise – quantum noise – which is due to spontaneous emission. One way of describing spontaneous emission is that it is emission "stimulated" by the zero–point quantum fluctuations of the electromagnetic field.

Achieving a spectral resolution of $R = \nu/\Delta\nu$ *requires that the instrument have some method of delaying the signal by a time* RT, or a distance $R\lambda$, where $T = \lambda/c$ is the period of the wave. This is a simple result of the Fourier transform relationship between frequency and time. The basic reason

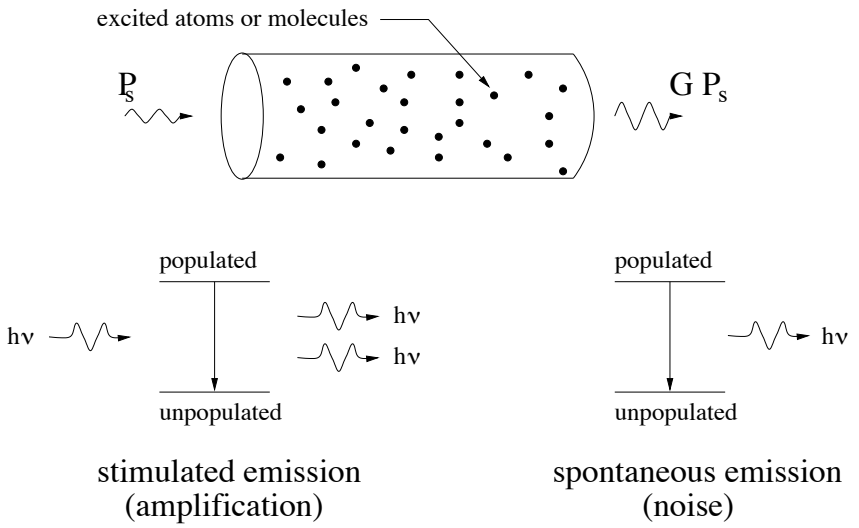

Figure 3: An illustration of quantum noise in a maser amplifier. This (fictitious) maser amplifier consists of a tube filled with a gas of molecules or atoms, which are pumped in a way that causes some transition with frequency ν to be inverted. A signal arriving at the input with power P_s is amplified by stimulated emission and emerges with power GP_s, where G is the power gain of the amplifier. However, due to spontaneous emission, noise photons emerge from the amplifier output even when $P_s = 0$.

that coherent detection is capable of very high spectral resolution, easily in excess of 10^6, is that the spectroscopy is actually done after downconversion, at radio or microwave frequencies, by the "backend". A wide variety of devices, such as filterbanks, acousto–optic (AOS) spectrometers, analog correlators, and digital correlators have been developed, which use various tricks to reduce the volume associated with the time delay RT. In essence, digital correlators store the digitized signal into memory for retrieval at a later time. Filterbanks and analog correlators use guided–wave (transmission line) propagation and dielectric materials to drastically reduce the volume. AOSs and CTSs (chirp–transform spectrometers) rely on the slow velocity of sound propagation in solids.

Another key point is that the *backend spectrometer does not need cryogenic cooling* because its noise is basically irrelevant, since the signal has been amplified. In contrast, the spectroscopic components of direct–detection instruments must be cold ($k_B T << h\nu$) to reduce the thermal background noise.

5. Direct Detection Spectroscopy

A wide variety of instruments have been used for direct detection spectroscopy: gratings, Fabry–Perots, Fourier Transform Spectrometers (FTS), etc. The choice depends on various requirements, including the wavelength range, the resolution and sensitivity required, the detectors available, the background level, size, cost, and so on. This plethora of approaches obscures the fact that *there is only one optimal approach that gives the best sensitivity: grating spectrometers or their equivalent.* Such spectrometers use a grating to disperse the light onto an array of detectors, and each detector pixel responds to a different wavelength channel.

In order to obtain maximum sensitivity, spectrometers should obey a simple principle: they must *extract the necessary information from every photon.* In a grating spectrometer, the absorption of a photon by a given detector pixel corresponds to a measurement of its wavelenth, to within the resolution of the instrument. This is not true for other types of spectrometers. For instance, the absorption of a photon by a detector in an FTS is not equivalent to a unique measurement of its wavelength. This corresponds to a loss of information, and therefore sensitivity. Another example is a Fabry–Perot spectrometer, which violates this principle by reflecting or "throwing away" photons outside of its bandwidth. The information carried by those photons is lost. Fabry–Perots must be scanned to obtain a spectrum, which is the time penalty that is paid for throwing away photons. The conclusion is that for best sensitivity, only grating spectrometers (or their equivalent) should be used.

An interesting corollary is that correlation spectroscopy, widely used in heterodyne backends, is a fundamentally inferior technique when used for direct detection. In contrast, it is well known that there is no sensitivity penalty for using correlators as backends. What is the difference ? Heterodyne backends measure classical signals, with very high photon occupation numbers, whereas direct–detection instruments on cold telescopes operate in the low occupation number regime.

In this context we note that the SPIRE instrument on Herschel will use an FTS for submillimeter spectroscopy. While there were numerous practical reasons for this choice, including limitations on the available detector sensitivity, the choice of an FTS means that SPIRE cannot achieve the maximum possible

sensitivity for spectroscopy. *A grating spectrometer, along with a much colder telescope, can offer spectacular sensitivities for spectroscopy*, many orders of magnitude better than SPIRE/Herschel. This is a very interesting possibility for the future, and is discussed in more detail in the presentations and papers by J. Bock, C. M. Bradford, and J. Glenn at this workshop.

The difficulty with grating spectrometers is that their size grows as the spectral resolution increases. For a resolution R, the linear size must be of order $R\lambda$, according to the time–delay principle described earlier. Achieving $R = 10^6$ at $\lambda = 200\,\mu$m, which can readily be done using a heterodyne spectrometer, would require a 200 m grating. Furthermore, this grating must be cold, to avoid a sensitivity degradatation. While there are no fundamental limitations, there are obviously enormous practical problems with this approach.

One is forced to look at ways of reducing the volume. Using guided–wave propagation helps; this is the key idea behind the waveguide grating spectrometer concept described by Bradford et al. at this workshop. Other ideas have been suggested, such as using the slower propagation in high–index materials such as silicon or germanium. However, none of these approaches is very likely get to $R = 10^6$.

The remaining possibility is to fold the optical path onto itself. This is exactly what is done in a Fabry–Perot, and does indeed give a large volume reduction, and can achieve resolutions approaching $R = 10^6$. The price is reduced sensitivity since photons are thrown away.

6. Comparison of coherent and direct detection

There are two ways to perform this comparison. One can look at demonstrated sensitivities of existing instruments, or careful design studies, including the limitations of existing detectors. In many ways this is the best approach, since it includes all of the "real life" factors – filter inefficiencies, excess noise, etc. This is the approach that should be used before starting construction of a major instrument. On the other hand, it is not helpful for projecting into the future.

The opposite extreme is to assume that fundamental limits will be approached closely. This is likely to be a very accurate projection for the future, but on an uncertain time scale. This is (almost) the approach that I will take, as shown in Figure 4. The one arbitrary "real–life" adjustment I have made

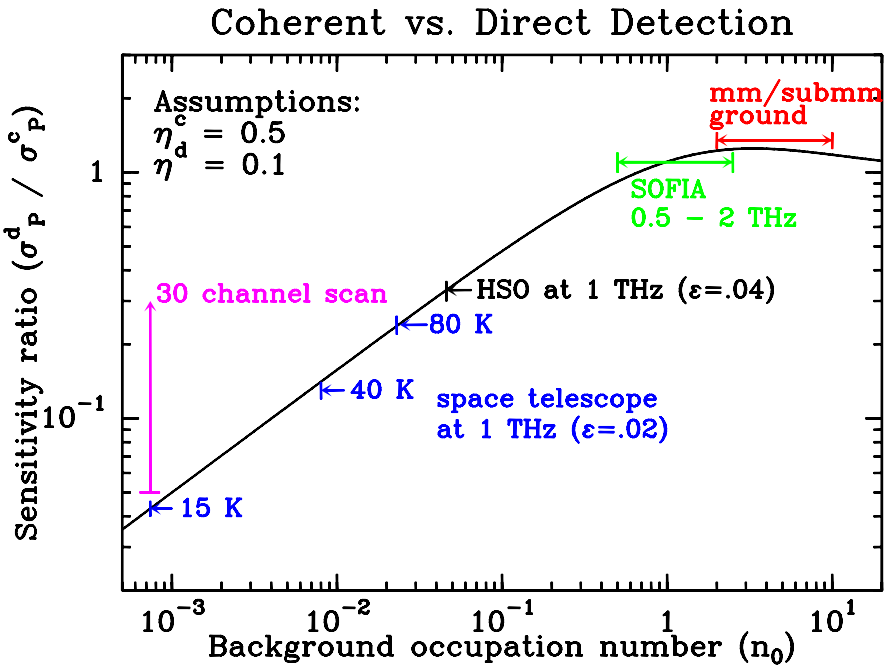

Figure 4: An idealized comparison of the relative sensitivities of coherent and direct detection. The vertical axis is the sensitivity ratio; a ratio less than unity favors direct detection. The vertical arrow shows the sensitivity penalty associated with a 30–channel sequential spectral scan. Sensitivity equations may be found in [1].

is to give the coherent instrument a better overall efficiency, since in reality it would have a much simpler optical system.

The sensitivity of ideal direct detection depends entirely on the background. This can be quantified by specifying the mean photon occupation number n. Ideal quantum–limited coherent detection corresponds to $n = 1$. For $n \sim 1$, there is no advantage to direct detection – which is the case for ground–based and airborne mm/submm/far-IR instruments. On the other hand, for $n << 1$, direct detection has an advantage which scales as $\sqrt{n}$, because coherent detection is quantum–limited with an equivalent background of $n = 1$.

For high spectral resolution, say $R \sim 10^6$, direct detection instruments

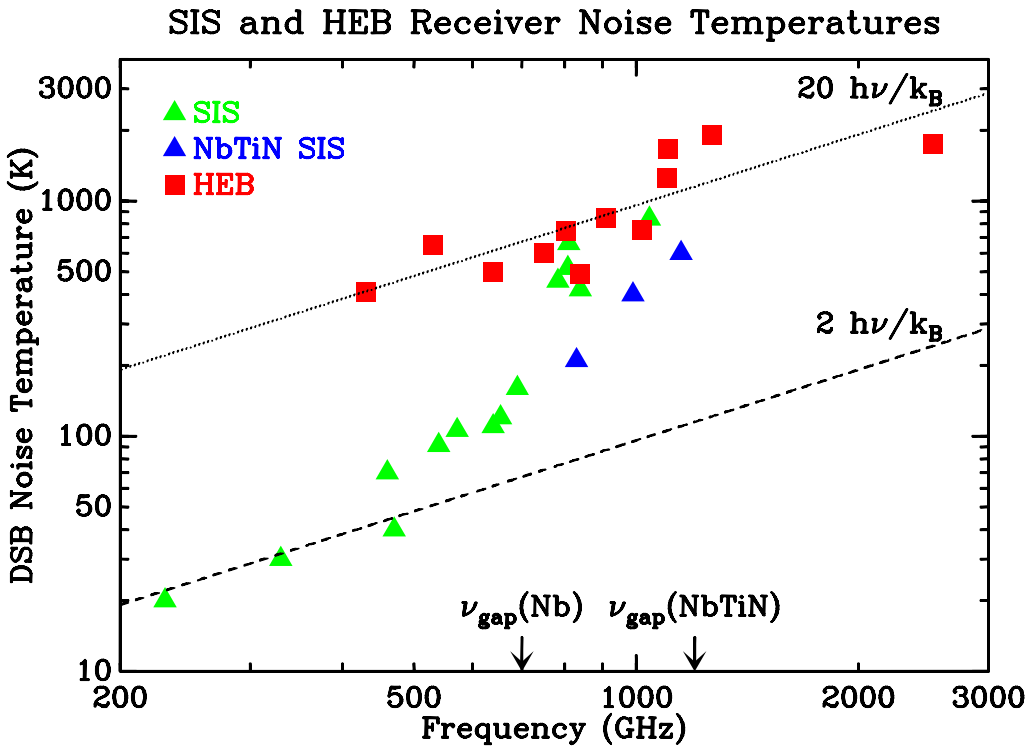

Figure 5: A selection of measured sensitivities for SIS and HEB receivers.

would use Fabry–Perots (as discussed in section 5) and would need to be scanned. For a scan consisting of M spectral channels, the sensitivity penalty would be $\sqrt{M}$, as shown by the vertical arrow on Fig. 5 for the case $M = 30$. The overall result is that direct detection is still more sensitive, but not by an overwhelming factor. However, there are several other factors to consider:

- backends can easily provide thousands of simultaneous channels
- backends can provide a wide range of spectral resolutions
- submillimeter heterodyne receivers are within a factor ~ 10 of the quantum limit (see Fig. 5)
- mixer noise temperatures degrade at higher frequencies
- tunable local oscillators are not yet available above ~ 1.5 THz

- background–limited $R \sim 10^6$ spectroscopy will require detectors that are $\sim 10^3$ times more sensitive than are now available ($10^{-21}\,\mathrm{W\,Hz^{-1/2}}$ vs. $10^{-18}\,\mathrm{W\,Hz^{-1/2}}$)

At present, these factors combine to strongly favor coherent detection for high–resolution spectroscopy up to $\sim 1.5-2$ THz. The current lack of tunable local oscillators limits the role of coherent detection at higher frequencies. In the future, we can expect the issue of detector sensitivity to disappear (for both direct and coherent). Even in this idealized case, it appears that coherent detection will retain substantial advantages in at least some situations, such as wideband, high–resolution line surveys.

7. Status of submm/far–IR receiver technology

Figure 5 shows the impressive improvements in receiver sensitivities that have been achieved over the last decade using superconducting mixers, with both tunnel junctions (SIS) and hot–electron bolometers (HEB). Nonetheless, there remains substantial room for improvement to reach the quantum limit, particularly at frequencies above 1 THz.

Local oscillators are another area in which dramatic improvements have been made. Electronically tunable, all solid–state local oscillators are being developed for Herschel. High–power transistor amplifiers at frequencies near 100 GHz are being used to drive diode multiplier chains to produce usable output power at frequencies as high as 1.5 THz.

Future developments can be expected in several areas. First, and most importantly, the push toward quantum–limited sensitivities must be continued. It has been demonstrated that SIS mixers are capable of reaching the quantum limit at millimeter wavelengths. However, SIS mixers become increasingly difficult to produce at higher frequencies, and will not operate above 1.5–1.6 THz with current materials. For higher frequencies, HEB mixers are used. Whether or not HEB mixers can ultimately reach the quantum limit is still an open issue; new device concepts may be required. Another area of development is to continue to expand the mixer instantaneous bandwidths. For ALMA, the goal is 8 GHz; work at Caltech is pushing toward 12 GHz. A third area, important for future space missions, is to look at integrating the later stages of the local oscillator with the mixer, in order to simplify the local oscillator

injection problem. Finally, there is still substantial room for development and improvement of local oscillators.

Acknowledgements

Submillimeter receiver development at Caltech is supported by NASA through the HIFI/Herschel, SOFIA, and ROSS/SARA NRA programs.

References

[1] J. Zmuidzinas, "Progress in coherent detection methods," in *The Physics and Chemistry of the Interstellar Medium* (V. Ossenkopf et al., ed.), pp. 423–430, U. Cologne, GCA–Verlag Herdecke, 1999.

[2] J. Fischer, "Galaxies: The long wavelength view," in *Proceedings of the Conference "ISO beyond the Peaks", Villafranca del Castillo, Spain, 2–4 February 2000 (ESA SP–456, 2000).*

[3] P. Schilke, D. J. Benford, T. R. Hunter, D. C. Lis, and T. G. Phillips, "A line survey of Orion–KL from 607 to 725 GHz," *Ap. J. (Suppl.)*, vol. 132, pp. 281–364, February 2001.

[4] J. Zmuidzinas, *Heterodyne Spectroscopy of Neutral and Ionized Carbon in the Interstellar Medium*. PhD thesis, University of California, Berkeley, 1987.

[5] M. L. N. Ashby *et al.*, "An analysis of water line profiles in star formation regions observed with the Submillimeter Wave Astronomy Satellite," *Ap. J. (Lett.)*, vol. 539, pp. L115–L118, August 2000.

[6] P. Schilke, D. M. Mehringer, and K. M. Menten, "A submillimeter HCN laser in IRC+10216," *Ap. J. (Lett.)*, vol. 528, pp. L37–L40, January 2000.

[7] N. Z. Scoville, M. S. Yun, and P. M. Bryant, "Arcsecond Imaging of CO Emission in the Nucleus of Arp 220," *Ap. J.*, vol. 484, pp. 702+, July 1997.

[8] K. Sakamoto, N. Z. Scoville, M. S. Yun, M. Crosas, R. Genzel, and L. J. Tacconi, "Counterrotating Nuclear Disks in Arp 220," *Ap. J.*, vol. 514, pp. 68–76, Mar. 1999.

[9] J. S. Ward, J. Zmuidzinas, A. I. Harris, and K. G. Isaak, "A ^{12}CO $J = 6 - 5$ Map of M82: The significance of warm molecular gas," *Ap. J.*, 2001. (submitted).

[10] C. M. Caves, "Quantum limits on noise in linear amplifiers," *Phys. Rev. D*, vol. 26, pp. 1817–1839, 1982.