Kölner Observatorium für MM- und Submm-Astronomie

Noise at Direct- and Heterodyne-Detection in the Infrared

by R. Schieder, KOSMA

Introduction

According to recent developments in heterodyne technology for the FIR and mid-IR frequency range the question should be raised, under which conditions this technology can compete with direct detection methods. Advantages and disadvantages of heterodyne and direct detection are frequently discussed in literature (see e.g. [1], [2], [3], [4]), but, due to the different technological aspects there is some lack in mutual understanding of the definitions and arguments so that a valid comparison of the methods remains still rather difficult. It is the purpose of this paper to present a detailed comparison, as well as to collect some arguments in favour of heterodyne methods in particular. Since several years heterodyne spectroscopy has also been introduced in the mid-infrared frequency range for astronomical observations at frequencies near 30 THz (i.e. λ=10µm), and the results are very promising (the GSFC/NASA receiver HIPWAC [5], the UC Berkeley Infrared Spatial Interferometer Array ISI [6], or the KOSMA Tuneable Heterodyne Infrared Spectrometer THIS [7]). With these instruments it is demonstrated that many scientific topics can be studied successfully, which are otherwise not so easily accessible (see e.g. [8], [9]). With the availability of tuneable Quantum Cascade Lasers (QCLs) heterodyne remote sensing throughout the mid-IR is now a valid option, and it becomes therefore more and more important to understand the characteristics of such instruments in this frequency range in comparison to direct detection instruments. Certainly, whether or not this method is useful and sensitive at fairly high frequencies is still debatable, and it needs a detailed analysis of all aspects involved before one can reach final conclusions.

When starting a discussion, it is mandatory to investigate all the details of eventually planned observations in order to establish a reasonable basis for a comparison. One of the most important parameters is the desired frequency resolution of the system. There are many objects of scientific interest, requiring very high frequency resolution like atomic or molecular spectra from cold areas in the interstellar medium, from comets, or from very cold and low pressure regions of planetary atmospheres. No doubt, a lot of information can be derived from high resolution spectra, but it is the question, what the required resolution really needs to be. For example, when studying lines originating from the interstellar medium in our galaxy, the typical velocity width is of the order of several km/sec so that a frequency resolution in the range of $R = \lambda/\delta\lambda \approx 10^5$ ($\delta V \approx 3$ km/sec) seems adequate in most cases, at least for the derivation of integrated line intensities. But, when studying details of line-shapes or narrow features of the cold gas spectra in quiescent interstellar clouds for example, higher resolution is desirable. Also, at temperatures below ambient, the Doppler width of molecular lines of heavier molecules needs a frequency resolution near 10^6 in order to resolve them sufficiently. This is particularly important for the analysis of atmospheric features. In the mid-infrared one has the advantage of much higher spatial resolution as compared to mm-, submm-, or far-IR observations. Consequently, atomic or molecular line features mostly tend to become narrower. In conclusion one can state that spectrometers with a resolution better than $R = 10^5$ are valuable instruments for many specific applications.

The question is, whether direct or heterodyne methods are better suited for such resolution. The standard direct detection spectrometer uses a grating as dispersive element. The required size of the grating scales with the product of resolution R and the wavelength λ. In consequence, at a resolution of 10^5 the size of the grating needs to be already of the order of 1 m when working at 10 µm wavelength. At sub-mm frequencies (500 GHz) it would increase to more than 50 m! Similar arguments apply to the size of the optics in Fourier-Transform machines as well as for a Fabry-

Perot type interferometer. It becomes therefore more and more unpractical to build direct detection spectrometers with very high resolution at longer wavelength. Despite of this, it is well known that such spectrometers suffer seriously from huge coupling losses when attempting very high frequency resolution, which has drastic influence on the sensitivity.

In comparison, the resolution of heterodyne receivers is achieved by electronic means so that any desired resolution is possible without the penalty of high losses in efficiency. On the other hand, heterodyne detectors have a fundamental limit in sensitivity caused by the existence of the so called "quantum limit", which does not apply for direct detectors. Due to this, direct detection can be significantly more sensitive. But this advantage disappears at longer wavelength due to the inevitable pick-up of thermal background at ambient temperature, which becomes more and more dominant below 6 THz ($h\cdot\nu \approx k_B\cdot T_A$). In conclusion one can state that heterodyne receivers generally become equivalent or even more sensitive than direct detection receivers in the longer wavelength regime. The turnover point is strongly dependent on frequency resolution and on the efficiencies of the hardware involved.

We concentrate in the following on observations of unresolved point-like sources. Therefore, single spatial mode coupling is analyzed. Since future applications point towards interferometers, heterodyne methods may become more interesting because of the well established delay-line technology and the ease determining the correlation function of pairs of telescopes. For simultaneous high frequency and very high spatial resolution it is therefore likely that heterodyne spectroscopy becomes an important tool also in the mid-infrared. On the following pages formulas are used, which are well documented in the literature, but a couple of modified arguments are included in order to support some of the conclusions. The different languages in the direct- and heterodyne detection world, i.e. the definitions of NEP and noise temperature in particular, need some cross-translation in order to make the sensitivity figures comparable. This is largely missing or misleading in the literature, and consequently some of the frequently used arguments in favor of one of the two methods are not quite applicable. Instead of dealing with signal to noise ratios, we concentrate on the formulation of "System Noise Temperature" (T_{Sys}) and/or "Noise Equivalent Power" (NEP), since it seems to deliver more insight into the matter. Therefore, the following discussion differs sometimes slightly from the standard treatments.

1. Direct Detection

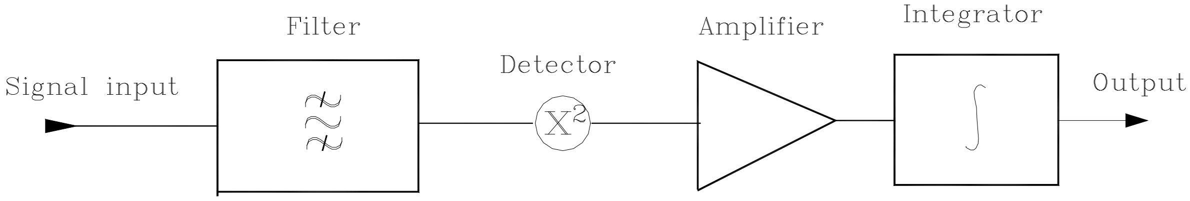

Fig.1: Scheme of a direct detector system with spectral resolution.

A sketch of one frequency pixel of a direct detection system with spectral resolution is shown in Fig.1. The resolution is passively done with a cooled filter in front of the detector. It is not intended to characterize all the various detectors in use, which all have different noise characteristics when considering the details. The simple assumption is that it provides the cycle averaged square of the input amplitude, and we just assume that there are charges generated which somehow contribute to the noise. We start with the uncertainty of a radiometer current, as can be extracted from the detector. With a given post-detection bandwidth Δ_{pd}, which is usually determined by the integration time t_{Int} of a final box-car integrator ($\Delta_{pd} = (2\cdot t_{Int})^{-1}$), we have for the current fluctuations of the detector signal:

$$<\Delta i^2> \; = \; 2\cdot[<i_{Fl}>^2/B_{Fl} + e_0\cdot<i_{Sh}>] \cdot \Delta_{pd} \; = \; <i_{Fl}>^2/(B_{Fl}\cdot t_{Int}) + e_0\cdot<i_{Sh}>/t_{Int}$$

The two terms in the brackets are of different origin: The first describes the so-called “wave fluctuation noise” or "radiometric noise", while the second represents the current induced shot noise. B_{Fl} is the “fluctuation bandwidth” of the filter, which is always larger than its resolution bandwidth δ_{Res}.[1] Only that part of the detector current contributes to the wave fluctuation noise, which is originates from frequency filtered photons. Usually, the wave fluctuation noise is not considered for direct detection instruments, but, since we are interested in high frequency resolution (B_{Fl} small), it is necessary to include it here. The corresponding photo current i_{Fl} consists now of a background current i_B and the signal current i_S.

The shot noise term i_{Sh} is comprised of the total current, which is due to photoelectrons originating from all filtered photons as well as due to charges caused by read-out noise, dark current or unfiltered photons. The latter contributions are summarized with the dark current i_D and are dependent on the type of detector. We rewrite now:

$$<\Delta i^2> \;=\; (<i_S>+<i_B>)^2/(B_{Fl}\cdot t_{Int}) \;+\; e_0\cdot(<i_S>+<i_B>+<i_D>)\,/\,t_{Int}$$

This is the detector contribution alone, but generally, post-detection amplification is needed which is responsible for additional noise. The amplifier we may describe by its noise temperature T_A, which characterizes the spectral noise power P (noise power per post-detection frequency interval) available at the output of the amplifier. It is referenced to the input of the amplifier when dividing by its effective power gain, which includes all eventual coupling losses between detector and amplifier. The effective contribution of amplifier noise to the system is equal to $4\cdot k_B\cdot T_A\cdot\Delta_{pd} = 2\cdot k_B\cdot T_A/t_{Int}$ (k_B is the Boltzmann constant). Thus we get now for the equivalent current fluctuations:

$$<\Delta i^2> \;=\; [\,(<i_S>+<i_B>)^2/B_{Fl} \;+\; e_0\cdot(<i_S>+<i_B>+<i_D>) + 2\cdot k_B\cdot T_A/R_A\,]\,/\,t_{Int} \qquad (1)$$

R_A is the effective impedance at the input of the amplifier. By definition, the minimum detectable signal current $<i_S^{Min}>$ is identical with the standard deviation of the current fluctuations $\sqrt{<\Delta i^2>}$, so that Eq.(1) represents a quadratic expression for $<i_S^{Min}>$, from which a general solution can readily be derived. In the limit of long integration time t_{Int} the result is:[2]

$$<i_S^{Min}> \;=\; [<i_B>\cdot(<i_B>/B_{Fl}+e_0) + e_0\cdot<i_D> + 2\cdot k_B\cdot T_A/R_A]^{1/2}\,/\,\sqrt{t_{Int}}$$

Assuming single spatial mode coupling, the background current $<i_B>$ can be converted into background spatial mode occupation numbers n_B (number of photons per second and Hz in one spatial mode) by:

$$<i_B> \;=\; e_0\cdot\int\; [\eta_\parallel\cdot n_{B\parallel}(\nu)+\eta_\perp\cdot n_{B\perp}(\nu)]\cdot L(\nu)\cdot d\nu = \; 2\cdot e_0\cdot\eta_D^{eff}\cdot\underline{n}_B\cdot\delta_{Res}$$

$\eta_\parallel$ and $\eta_\perp$ are the coupling efficiencies of the instrument to the parallel (in an arbitrary orientation) and orthogonal polarization of the incident radiation field. They include the throughput of the filter at maximum as well as the quantum efficiency of the detector for both polarizations and are treated as constant within the filter width. Since the η_i are eventually different, like is true for grating spectrometers for example, an effective efficiency $\eta_D^{eff} = ½\cdot(\eta_\parallel+\eta_\perp)$ is introduced. $n_{B\parallel}(\nu)$ and $n_{B\perp}(\nu)$ are the single mode occupation numbers for the two orthogonal polarizations of the radiation. We assume that they are identical, as is true for thermal radiation emitted from an isotropic surface. We

[1] For the mathematical definitions see e.g. in [20]:

$$\delta_{Res} = \int L(\nu)\cdot d\nu \qquad B_{Fl} = [\int L(\nu)\cdot d\nu]^2 / \int L^2(\nu)\cdot d\nu$$

L is the normalized power transmission of the filter versus frequency ν with $0 \le L(\nu) \le L_{Max} = 1$. In the following treatment the ratio q of these two quantities is used, which is

$$q \;=\; B_{Fl}\,/\,\delta_{Res} \;=\; \int L(\nu)\cdot d\nu\,/\int L^2(\nu)\cdot d\nu.$$

q is therefore always larger or equal to 1. For a grating spectrometer at highest resolution its value is exactly 1.5, when evaluating the formulas above. For a high resolution Fabry-Perot interferometer it reaches a value of 2, while it becomes unity for a filter with rectangular shape. For practically all realistic filters one finds: $1 < q \le 2$.

[2] At zero background and zero detector dark current $<i_d>$ the solution looks different. In this case we have:

$$<i_S^{Min}> \;=\; e_0/t_{Int}$$

This corresponds to the detection of one electron per observing period. On the other hand, at very short integration time near to or smaller than $1/B_{Fl}$, the contribution of the wave fluctuation noise is no more described by the standard radiometer equation, and Eq.(1) is not valid anymore. A detailed discussion can be found for example in [14].

also assume that they are constant within the filter width, therefore we have set $n_{B\parallel}(v) = n_{B\perp}(v) = \underline{n}_B$. Frequently, more than one single spatial mode may contribute to the background. In this case $\underline{n}_B$ must be replaced by $s{\cdot}\underline{n}_B$ with s the effective number of background spatial modes as seen by the detector.[3] For a system with high spatial resolution s should not be much larger than 1.

The minimum signal photo current i_S^{Min} can be converted into the minimum incident signal power P_S^{Min} by using

$$<i_S^{Min}> \ = \ \eta_D \cdot P_S^{Min}/hv \cdot e_0$$

with h Planck's constant. v is the center frequency of the filter, and we have again assumed that the filter width is small. The signal detection efficiency η_D is equal to η_D^{eff} for un-polarized and equal to one of the η_i for polarized signal radiation. We have now:

$$P_S^{Min} \ = \ hv/\eta_D{\cdot}\delta_{Res}{\cdot}\{2{\cdot}\eta_D^{eff}{\cdot}s{\cdot}\underline{n}_B{\cdot}(2{\cdot}\eta_D^{eff}{\cdot}s{\cdot}\underline{n}_B+q)+[<i_D>/e_0+2{\cdot}k_B{\cdot}T_{pd}/(e_0^2{\cdot}R_A)]\ {\cdot}q/\delta_{Res}\}^{1/2} / \sqrt{}(B_{Fl}{\cdot}t_{Int}) \quad (2)$$

The efficiency η_D is usually fairly small for high resolution spectrometers due to low throughput so that one has substantial loss in sensitivity when attempting very high frequency resolution. The background radiation is not polarized, therefore $2{\cdot}\eta_D^{eff} \ = \ \eta_\parallel+\eta_\perp$ stands for its contribution. For thermal radiation the single spatial mode occupation number of the background $\underline{n}_B$ is given by Bose-Einstein:

$$\underline{n}_B \ = \ f/[e^{T_Q/T} - 1], \quad T_Q = hv/k_B \qquad (3)$$

f is the emissivity of the involved surfaces at physical temperature T.

When neglecting all detector and amplifier contributions, Eq.(2) corresponds exactly to the relation as found in [10] for example with one small discrepancy: q has there a value of unity, which means that no distinction between resolution and fluctuation bandwidth is made. Only a box-car shaped filter could justify this, which is probably not very realistic. At small background occupation numbers only q survives in the brackets, i.e. the system is purely shot noise limited and the minimum detectable power is proportional to the square root of the background power. On the other hand, if the background becomes large, we end with the well established radiometer equation [13].

The *NEP* [4] is defined as the minimum detectable power of the system for a post-detection bandwidth of 1 Hz, which is equivalent to an integration time of 0.5 seconds. Therefore we have:

$$\begin{aligned} NEP \ &= \ P_S^{Min}(t_{Int}= 0.5\ sec) \ = \\ &= \ \sqrt{}2/\eta_D \cdot k_B{\cdot}T_Q \cdot \{\ \delta_{Res}/q \cdot 2{\cdot}\eta_D^{eff}{\cdot}s{\cdot}\underline{n}_B \cdot (2{\cdot}\eta_D^{eff}{\cdot}s{\cdot}\underline{n}_B+q) + <i_D>/e_0 + 2{\cdot}k_B{\cdot}T_{pd}/(e_0^2{\cdot}R_A)\ \}^{1/2} \qquad (4) \end{aligned}$$

The units of the *NEP* are Watts/Hz$^{1/2}$, and one should keep in mind that the root Hz refers to the post-detection and not to the pre-detection bandwidth. It should be mentioned that there exist different definitions of the *NEP* in the literature. Sometimes, an additional factor of $\sqrt{}2$ is applied, because a complete observation is the difference of two independent measurements, a "signal" and a "reference" measurement of equal length (t_{Int} = 0.5 sec each, t_{total} = 1 sec). We do not use it here in order to make the *NEP* and T_{Sys} more equivalent.

For thermal background we may replace the expression $s{\cdot}\underline{n}_B$ by the total background power P_B with

$$2\ {\cdot}s{\cdot}\underline{n}_B \cdot \delta_{Res} = \ P_B/hv$$

v is again the mean frequency of the filter. Then, we can rewrite the *NEP* in terms of the background power P_B:

[3] This is a somewhat simplistic approach and deserves a more detailed analysis. We can interpret s as the effective mode number which accounts for all eventual correlation effects between the modes. See for example in [12].

[4] Frequently the "Noise Equivalent Flux Density" (*NEFD*) is introduced for the characterization of spectrometers. It is defined as minimum detectable power per frequency interval, telescope area and root of the post-detection bandwidth. Therefore we have:

$$NEFD \ = \ NEP\,/\,(A_{eff} \cdot \delta_{Res}) \quad \text{[spectral intensity/Hz}^{1/2}\text{]}$$

A_{eff} is the effective telescope area, which takes the sensitivity distribution across the telescope surface area into account. The *NEFD* is usually given in Jy/Hz$^{1/2}$, while 1 Jy = $1{\cdot}10^{-26}$ Watt/(Hz·m^2). (The "Hz" refers here to the pre-detection bandwidth δ_{Res}.)

$$NEP = \sqrt{}2/\eta_D \cdot k_B \cdot T_Q \cdot \{ \eta_D{}^{eff} \cdot P_B/h\nu + (\eta_D{}^{eff} \cdot P_B/h\nu)^2/B_{Fl} + <i_D>/e_0 + 2 \cdot k_B \cdot T_{pd}/(e_0{}^2 \cdot R_A) \}^{1/2}$$

One can see here that the wave fluctuation noise is irrelevant as long as the rate of detected background photons is much smaller than the fluctuation bandwidth B_{Fl}. In other words, the spectral density of the background is decisive for a contribution of the fluctuation noise. When dealing with this formula one should keep in mind that the background power P_B is proportional to the resolution bandwidth δ_{Res}. The influence of background radiation is therefore easier to estimate when using Eq.(4).

In heterodyne spectroscopy, the system noise temperature is generally introduced to characterize the noise performance of a receiver, and in many cases it may be a useful expression for direct detection systems as well. It is not surprising that Eq.(2) has the same mathematical appearance as the radiometer formula, which is the standard description of the noise in a heterodyne system (see below). If we consider $P_S{}^{Min}$ as the standard deviation of the detected power fluctuations, we can write according to Eq.(2):

$$P_S{}^{Min} = \Delta P_{Sys} = P_{Sys} / \sqrt{}(B_{Fl} \cdot t_{Int}) \text{ with}$$

$$P_{Sys} = 1/\eta_D \cdot k_B \cdot T_Q \cdot \delta_{Res} \cdot \{2 \cdot \eta_D{}^{eff} \cdot s \cdot \underline{n}_B \cdot (2 \cdot \eta_D{}^{eff} \cdot s \cdot \underline{n}_B + q) + [<i_D>/e_0 + 2 \cdot k_B \cdot T_{pd}/(e_0{}^2 \cdot R_A)] \cdot q/\delta_{Res} \}^{1/2}$$

P_{Sys} is the noise power generated by the instrument itself, and, if we divide by the bandwidth δ_{Res}, we get the spectral power density of the noise, and, when further dividing by the Boltzmann constant, we derive an expression for the so called "system noise temperature":

$$T_{Sys} = P_{Sys}/(\delta_{Res} \cdot k_B)$$

Generally, the system noise temperature is defined for the spectral power density in a single spatial mode. Therefore we need to distinguish between single (SP) and dual polarization (DP) observations. For a dual polarization situation, we deal with two spatial modes, one for each polarization. In this case the "dual polarization" system temperature $T_{Sys}(DP)$, which describes the contribution per polarization, is reduced by the sum of the efficiencies of both polarizations.

$$T_{Sys}(DP) = T_Q/(\eta_\parallel + \eta_\perp) \cdot \{2 \cdot \eta_D{}^{eff} \cdot s \cdot \underline{n}_B \cdot (2 \cdot \eta_D{}^{eff} \cdot s \cdot \underline{n}_B + q) + [<i_D>/e_0 + 2 \cdot k_B \cdot T_{pd}/(e_0{}^2 \cdot R_A)] \cdot q/\delta_{Res} \}^{1/2} \qquad (5)$$

The expression shows that the system noise temperature decreases with increasing resolution bandwidth δ_{Res} as long as the system is not background limited. The dual and the single polarization system noise temperatures $T_{Sys}(DP)$ and $T_{Sys}(SP)_i$ are related by

$$T_{Sys}(SP)_i = (\eta_\parallel + \eta_\perp)/\eta_i \cdot T_{Sys}(DP) = 2 \cdot \eta_D{}^{eff}/\eta_i \cdot T_{Sys}(DP), \quad i = \parallel, \perp$$

If the efficiencies η_i of both polarizations are identical, the SP-system temperature becomes twice the DP-system temperature.

When comparing with the *NEP* (Eq.(4)), we have now:

$$T_{Sys}(DP) = 2^{-3/2} \cdot NEP(DP)/k_B \cdot [q/\delta_{Res}]^{1/2} \quad \text{and} \quad T_{Sys}(SP)_i = 2^{-1/2} \cdot NEP(SP)_i/k_B \cdot [q/\delta_{Res}]^{1/2}$$

Note that the dual polarization and single polarization *NEP*s are identical for identical $\eta_\parallel$ and $\eta_\perp$, while this is not the case for T_{Sys}. Note also that in case of background limited operation, i.e. at negligible amplifier noise and detector dark current, the *NEP* is proportional to the square root of the resolution bandwidth δ_{Res} while the system noise temperature is independent on δ_{Res}. This makes T_{Sys} a convenient definition for all background limited instruments. On the other hand, if the performance is determined by the detector/amplifier properties alone, the system noise temperature becomes inverse proportional to the square root of the resolution bandwidth while the *NEP* becomes independent on resolution. This makes the *NEP* now particularly useful. The overall system parameters determine which of these two descriptions looks more favorable.

The difficulty with the system noise temperature of a direct detection system is that the brightness temperature of an external source does not co-add linearly to the noise temperature when dealing with strong signals. Only at very high background levels, where shot noise and amplifier noise can be neglected, the situation is different, and it is obvious, that the noise temperature definition becomes very handy in this case. At low background one has to co-add the background photon numbers to the photons from the source and use it as new “background” mode occupation number in Eqs.(4) and (5). This behavior is very different from that of heterodyne receivers (see below).

1.a. Estimates of Direct Detection Sensitivity

The sensitivities of direct detection spectrometers in the FIR and the mid-IR are determined by rather different parameters. This is mainly caused by the different contributions of the background radiation to the system temperature. Due to Bose-Einstein, the brightness temperature J of a black body source at temperature T is given by [5]:

$$J \;=\; T_Q/(e^{T_Q/T}-1) \qquad (6)$$

Thus, an ambient temperature emitter at 295 K contributes a signal equivalent to a brightness temperature J of 283 K at 500 GHz, which is nearly a factor of 12 more than the quantum limit at that frequency (T_Q = 24 K). In contrast, at 30 THz the corresponding brightness temperature is only 11 K equivalent to 0.8% of the quantum limit (T_Q = 1440 K). It is therefore crucial to consider the background contributions very carefully. The lower frequency range can gain drastically when going into space while cooling the telescope to Helium-temperatures. This will not help much in the mid-IR as long as the resolution bandwidth of the spectrometer is small. In order to analyze the situation in more detail, we consider first the mid-IR case.

At 10 µm wavelength, the mode occupation numbers are small. At ambient temperature we have values of $7.6\cdot10^{-3}$ photons/(sec·Hz) for each polarization component, which reduces to $\underline{n}_B$ = $7.6\cdot10^{-4}$ (see Eq.(3)), when assuming an emissivity of the involved surfaces of ε=0.1. This means that the shot noise, and not the wave fluctuation noise, is dominating the background contribution to the total noise. When considering the ideal case of zero detector and amplifier contribution, we find a dual polarization system temperature of 184 K (using Eq.(5) with q = 1.5 and s = 1), while assuming a typical effective efficiency of the system of $\eta_D^{eff} \approx 0.035$ at 10 µm wavelength and a resolution bandwidth of 300 MHz (R = 10^5), as is reported for the TEXES instrument [15]. Thus, the system noise temperature of a direct detection system can become significantly smaller than the quantum limit.

At the presently highest available resolution of $R = \nu/\delta_{Res} \approx 10^5$ TEXES should have a NEP of $1.0\cdot10^{-16}$ Watt/Hz$^{1/2}$ (see Eq.(4)). On the other hand, the reported $NEFD$ is about 13 Jy within 1 second on-source integration time (telescope radius = 1.5m, λ=10µ). This corresponds to a NEP of about $3.9\cdot10^{-16}$ Watt/Hz$^{1/2}$ or 700 K dual polarization noise temperature. Thus, the measured system temperature is not too far below the quantum limit (49%). The difference between theory and experiment may be due to the coupling of background radiation from more than one spatial mode, as was mentioned above. The number of background modes needs to be raised to $s \approx 14$ to match the experimental result. It appears therefore rather crucial to optimize spatial filtering in the cold part of the instrument.

In general, dark current and amplifier noise may play a significant role as well. They must be kept as small as possible. But, as is visible in Eqs.(4) and (5), their influence diminishes at low resolution, while the background radiation begins to dominate. The influence of the amplifier noise temperature is also of concern, unless the impedances of amplifier and detector are extremely high. For example, the amplifier noise temperature may be as low as 10 K. If the impedance is 10 GΩ, the calculated dual polarization noise temperature increases to 1510 K, which is nearly identical with the quantum limit. Therefore, impedances of more than 10^{11} Ω are required, which is not out of reach at very low temperatures. A calculation for the influence of detector dark current leads to similar conclusions. This demonstrates how important the detector parameters are when attempting high frequency resolution with a direct detection system.

[5] Note that the meaning of the phrase "brightness temperature" J is not that of a real temperature. The product $k_B\cdot J$ describes the spectral power density (Watt/Hz) as is emitted by a source into one spatial mode. Frequently, J is also considered as the physical temperature a thermal source would have, if the Rayleigh-Jeans approximation would be valid. This we consider as a rather artificial description, because it becomes fairly useless, in case non-thermal radiation is discussed.

At longer wavelength, e.g. 0.6 mm (500 GHz), the situation changes, because the thermal background starts to dominate. There we have a quantum limit at T_Q = 24 K, and, when using similar data as for the 10 µm case, the background contribution to the dual polarization noise temperature increases to 10 times the quantum limit. Therefore, at fixed absolute frequency resolution – i.e. at fixed size of the grating in the spectrometer – the situation becomes more and more background limited with increasing wavelength, or in other words: the dark current and amplifier contributions are not that important at longer wavelength, at least for instruments which are operated at ground-based observatories. This makes heterodyne instruments particularly competitive in this frequency range.

2. Heterodyne Detection

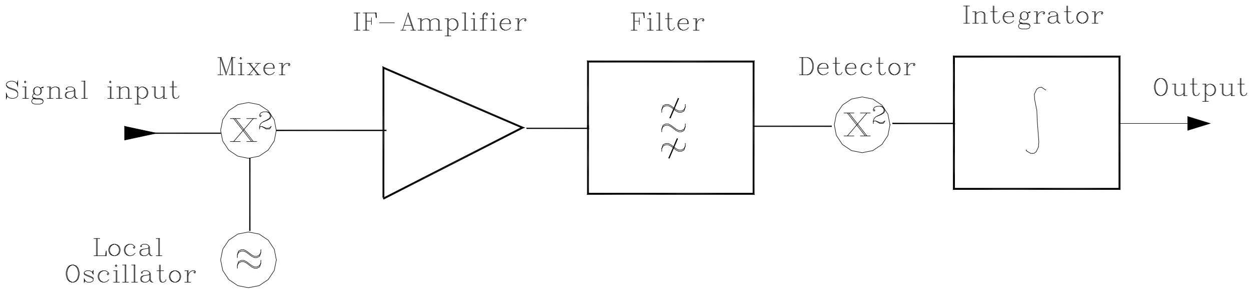

Fig.2: Scheme of a heterodyne receiver with spectral resolution.

When comparing direct detection and heterodyne receivers, one must be certain that one compares them correctly. Just consider both instruments as black boxes with an input port, where the signal radiation goes in, and an output port supplying a current, which changes proportionally to the input signal power. For a direct detector this introduces no complication, the signal to noise is directly determined by the electrical power of the detector signal current and the variance of its current fluctuations. For a heterodyne system one has to consider the current as is supplied by the quadratic detector in the backend. There are therefore two quadratic devices involved, the mixer detector and the backend detector.

For a more detailed comparison we have to investigate the various noise terms in a heterodyne system as well. A scheme of a heterodyne system is shown in Fig.2. Usually, one has to consider dual sideband reception of the mixer. In essence, this is more or less equivalent to the dual polarization response of a direct detector. In addition, there is now the "quantum limit" involved, which reduces the sensitivity of the receiver. It is also important to note that a heterodyne mixer is generally sensitive to one spatial mode per sideband and to one polarization only. And finally, the detection of the signal follows after significant amplification of the mixer signal so that the final detection in the backend is purely wave-fluctuation noise limited.

To keep it simple, let us start the discussion with a single sideband mixer, which is sensitive in one of the two possible sidebands only. The number of intermediate frequency photons (IF-photons) per frequency interval and time as generated by the mixer is given by:

$$n'(\nu_{IF}) \;=\; G_M \cdot \{\, 1+\eta_H \cdot n(\nu_S)\} - 1 \qquad (7)$$

ν_S is the signal frequency, and η_H is the efficiency of the mixer, which we assume as constant within the resolution bandwidth of the spectrometer. For a heterodyne mixer the sideband efficiency η_H may have values close to unity, since there is no filter in front with eventually poor throughput like in a direct detection system. ν_{IF} is the intermediate frequency, $n(\nu_{IF})$ is the number of IF-photons, while $n(\nu_S)$ is the mode occupation number in the signal band. G_M is the IF photon gain, which represents the number of IF-photons generated by each detected signal photon. Usually, there is photon gain G_M » 1 and it can be very large. For instance, a typical mm- or submm-Schottky mixer has a power conversion gain in the range of g_M = -10 dB or even better. (g_M is the ratio of absorbed signal power and generated IF-power.) The corresponding photon gain G_M is calculated by multiplying the power gain with the ratio of photon energies of signal and IF-photons, which is roughly 500 for a signal frequency of 500 GHz and an IF-frequency of 1 GHz for example. Thus, the photon gain of the mixer is near 50. If there is such gain, i.e. stimulated emission of IF-photons, then spontaneous emission must also occur. Therefore, there is always one IF-photon created in the mixer in each spatial mode per frequency interval and time, independent on the presence of signal photons. This is included by the additional “1” in the brackets of the equation above, since we assume only 1 spatial mode in the IF-circuitry. This description is equivalent to the quantum-mechanical treatment of spontaneous and stimulated emission of atoms or molecules. The “-1” at the end finally removes the contribution of the spontaneous emission in case there is no gain (G_M=1), and, if the gain G_M is large enough, we may neglect this term. (Similar arguments one can find for example in [16].) By the way, the description of the minimum noise of linear amplifiers (see e.g. [17], or [18]) is analogous.

There is also a shot-noise term to be included which originates from the current through the mixer. The spectral distribution of the current fluctuations is given by $2 \cdot e_0 \cdot \langle i_M \rangle$ with $\langle i_M \rangle$ the total mixer current. Therefore the spectral density of the available power is given by $2 \cdot e_0 \cdot \langle i_M \rangle \cdot R_M$. R_M is the dynamic IF-resistance of the mixer, and it can vary significantly with different mixers. Ideally, it should be near 50 Ω for a good match to an IF-amplifier. The number of additional IF noise photons per unit time and frequency is determined by the available power divided by the IF-photon energy. We split now the mixer current $\langle i_M \rangle$ into two parts, the photo current $\langle i_{LO} \rangle$, which is generated by the absorbed local oscillator power, and the mixer "bias" current $\langle i_B \rangle$ including all contributions which are not caused by absorption of LO photons. An eventual contribution due to a signal we neglect, since we investigate small signal response only. This leads to:

$$n'(\nu_{IF}) \;=\; G_M \cdot \{\, 1+\eta_H \cdot n(\nu_S) \,\} + 2 \cdot e_0 \cdot (c \cdot \langle i_{LO} \rangle + \langle i_B \rangle) \cdot R_M / h\nu_{IF} \qquad (8)$$

The empirical constant c describes how the photo-current contributes to the IF noise. For example, the LO photo-current in an ideal SIS mixer does not generate any shot noise [19]! This means c is zero, and, in case there is no external bias current, the total shot noise term becomes exactly zero. With a classical photo-detector with square law characteristics c becomes unity (see below). It therefore depends on the mixer type and the detailed conditions the mixer is operated, how the shot noise contributes to the total noise.

Behind the mixer an IF-amplifier is needed which we characterize by its net gain G_A. (G_A includes all eventual coupling losses between mixer and amplifier.) If we consider the amplifier as noiseless, we can write for the mode occupation number $n(\nu_{IF})$ behind the amplifier with G_M large:

$$n(\nu_{IF}) \;=\; G_A \cdot [1 + n'(\nu_{IF})] - 1 \;\approx\; G_A \cdot G_M \cdot \{\, 1 + \eta_H \cdot n(\nu_S) + 1/G_M \cdot 2 \cdot e_0 \cdot \langle i_M \rangle \cdot R_M / h\nu_{IF} \,\}$$

The power gain g_M and the IF photon number gain G_M are related by

$$G_M \cdot \nu_{IF} \;=\; g_M \cdot \nu_S$$

and we have now:

$$n(\nu_{IF}) \;=\; G_A \cdot G_M \cdot \{\, 1 + \eta_H \cdot n(\nu_S) + 1/g_M \cdot 2 \cdot e_0 \cdot \langle i_M \rangle \cdot R_M / h\nu_S \,\}$$

The IF-signal is then passed through a filter and is finally detected by a square-law detector in the backend. The efficiencies of the detector as well as the throughput of the filter are not important, because the number of IF-photons after amplification is very large. Any eventual loss can be considered as part of the effective IF-gain G_A.

The complete detector current in the backend can now be calculated with:

$$\begin{aligned} \langle I_{IF} \rangle \;&=\; b \cdot e_0 \cdot \int n(\nu_{IF}) \cdot L(\nu_{IF}) \cdot d\nu_{IF} \;\approx\; b \cdot e_0 \cdot \delta_{Res} \cdot \underline{n}(\nu_{IF}) \\ &=\; b \cdot e_0 \cdot \delta_{Res} \cdot G_A \cdot G_M \cdot \{1 + \eta_H \cdot \underline{n}(\nu_S) + 1/g_M \cdot 2 \cdot e_0 \cdot \langle i_M \rangle \cdot R_M / h\nu_S \,\} \end{aligned}$$

The constant b describes the efficiency for the conversion of IF-photons into detector electrons, but its value is unimportant here again. δ_{Res} is the resolution bandwidth of the filters in the backend, and $\underline{n}(\nu_{IF})$ is the average IF-mode occupation number within the width of the filter. (We have assumed again that the frequency distribution of the IF-signal is much broader than the filter width.) We now refer to the effective mixer current and divide $<I_{IF}>$ by the total gain $b \cdot G_A \cdot G_M$ and consider $I_{IF}/(b \cdot G_A \cdot G_M)$ as the equivalent mixer current i_{IF}:

$$<i_{IF}> \; = \; e_0 \cdot \delta_{Res} \cdot \{1 + \eta_H \cdot \underline{n}(\nu_S) + 1/g_M \cdot 2 \cdot e_0 \cdot <i_M> \cdot R_M / h\nu_S\} \qquad (9)$$

With large $n(\nu_{IF})$ the detected signal is only limited by wave fluctuation noise, whereas detector shot-noise in the backend becomes irrelevant. Therefore, the radiometer equation fully applies and we have for the noise:

$$<\Delta i_{IF}^{\;2}> \; = \; <i_{IF}>^2 / (B_{Fl} \cdot t_{Int})$$

B_{Fl} is the fluctuation bandwidth of the backend filter.

The minimum detectable equivalent current $<i_F^{\;Min}>$ is equal to the square root of $<\Delta i_{IF}^{\;2}>$. We introduce now the equivalent minimum detectable input power $\underline{P}_S^{\;Min}$ by setting:

$$<i_{IF}^{\;Min}> \; = \; e_0 \cdot \int \eta_H \cdot n_S^{\;Min}(\nu_S) \cdot L(\nu_{IF}) \cdot d\nu_{IF} \; \approx \; e_0 \cdot \eta_H \cdot \underline{n}_S^{\;Min} \cdot \delta_{Res} = e_0 \cdot \eta_H \cdot \underline{P}_S^{\;Min} / h\nu_S, \text{ with } \; \nu_S = \nu_{LO} + \nu_{IF}$$

$n_S^{\;Min}(\nu_S)$ is the incident minimum occupation number at the input side of the system, and $\underline{n}_S^{\;Min}$ is the average number within the full width of the filter, while we have used: $\underline{P}_S = \underline{n}_S \cdot h \cdot \nu_S \cdot \delta_{Res}$.

At very low signal levels the mode occupation numbers in Eq.(9) are those of the background photon flux, and we replace $n(\nu_S)$ by $n_B(\nu_S)$. We need to consider the integrated signal power $\underline{P}_S^{\;Min}$ as seen with the detector, and we get in the small signal limit:

$$\begin{aligned} \underline{P}_S^{\;Min} \; &= \; h\nu_S / \eta_H \cdot <i_{IF}^{\;Min}> / e_0 \; = \\ &= \; k_B \cdot T_Q / \eta_H \cdot \delta_{Res} \cdot \{1 + \eta_H \cdot n_B(\nu_S) + 1/g_M \cdot 2 \cdot e_0 \cdot <i_M> \cdot R_M / h\nu_S\} \, / \sqrt{(B_{Fl} \cdot t_{Int})}, \;\; T_Q = h\nu_S / k_B \qquad (10) \end{aligned}$$

$n_B(\nu_S)$ is for thermal radiation determined again by the Bose-Einstein relation modified by the emissivity factor f (see Eq.(3)). Note that the different noise sources add now linearly, whereas they combine as the root of the sum of the squares for direct detectors at low background contributions. This is the immediate consequence of the fluctuation noise behavior of a heterodyne system.

For the motivation of the introduction of a "system temperature" we can use the following arguments: Consider a thermal load, which emits power in each sideband within a given bandwidth δ_{Res} like:

$$P \; = \; k_B \cdot J \cdot \delta_{Res},$$

J is the “brightness temperature” of a single spatial mode, which might be described by the Bose-Einstein formula in case it is a thermal (black) emitter.[6] For the definition of the noise temperature we artificially consider the receiver as noiseless with a fully coupled broadband source in front. It emits with a certain brightness temperature T_{Mix}[7], which gives rise to the same fluctuations as the system itself generates. (We use here T_{Mix} instead of J_{Mix} because of the historical tradition.). Its total emitted power P_{Source} is for a small resolution bandwidth δ_{Res} given by:

$$P_{Source} \; = \; k_B \cdot T_{Mix}(SSB) \cdot \delta_{Res}.$$

$T_{Mix}(SSB)$ is the “single sideband system noise temperature” of the mixer. A change ΔP of the input power, i.e. a change of the brightness temperature T_{Mix} by ΔJ is detectable if

$$\Delta P \; = \; k_B \cdot \Delta J \cdot \delta_{Res} \; = \; P_S^{\;Min}$$

Thus we have for single sideband reception, when using Eq.(10):

$$\Delta J \; = \; 1/\eta_H \cdot T_Q \cdot \{\, 1 + \eta_H \cdot n_B(\nu_S) + 1/g_M \cdot 2 \cdot e_0 \cdot <i_M> \cdot R_M / (k_B \cdot T_Q) \,\} \, / \sqrt{(B_{Fl} \cdot t_{Int})}$$

(Note that this describes the sensitivity of the mixer alone and does not include any noise contributions from the IF-amplifiers.) When comparing this with the radiometer equation

$$\Delta J \; = \; T_{Mix} / \sqrt{(B_{Fl} \cdot t_{Int})}$$

we can define a single sideband system noise temperature for the upper or lower sideband with

[6] In general there is no need to introduce a Planck emitter here. Any other spectral power distribution might be useful as well. For example, the spectral power $k_B \cdot J$ may be generated by a non-thermal process like Bremsstrahlung, which happens to have the “brightness temperature” J at the frequency the receiver is sensitive to.

[7] In fact, the “noise temperature” should better be called “noise brightness temperature” since it characterizes the emitted spectral power $k_B \cdot T_{Mix}$ of the source which generates the same power fluctuations $k_B \cdot \Delta J$ as the receiver. It therefore assumes that the noise power fluctuations are proportional to the mean spectral power of the source.

$$T_{Mix}(SSB)_{u,l} \ = \ 1/\eta_{u,l} \cdot T_Q \cdot \{ \ 1 + \eta_{u,l} \cdot n_B(\nu_{u,l}) + 1/g_M \cdot 2 \cdot e_0 \cdot <i_M> \cdot R_M/(k_B \cdot T_Q) \ \} \qquad (8)$$

$\eta_{u,l}$ is the efficiency η_H for the upper or lower sideband respectively. $T_{Mix}(SSB)_{u,l}$ is the single sideband system noise temperature, which is analogous to the single polarization noise temperature before. Remember that this formula applies for mixers, which are sensitive in one sideband only!

Until now we have treated the noise generated by the mixer alone, while we have neglected the contribution of the following amplifier to the noise of the complete system. When including it we can use the well established expression (see e.g. [20]):

$$T_{Sys} = \ T_{Mix} + 1/\gamma \cdot T_{IF}$$

T_{IF} is the noise temperature of the IF-system and γ is the total mixer power conversion gain. We have therefore:

$$\gamma = P_{IF}/P_S(\nu_S)$$

$P_S(\nu_S)$ is the incoming signal power in the sideband and P_{IF} the IF-power delivered by the mixer. The gain g_M, as was introduced earlier, was defined as the ratio of generated IF-power and absorbed signal power. It did therefore not include optical losses between receiver input and mixer as well as the quantum efficiency of the mixer, which we describe together by the optical efficiency η_S. Thus we can write:

$$\gamma = \eta_S \cdot g_M$$

When combining all this together we have finally for the noise temperature:

$$T_{Sys}(SSB)_{u,l} \ = \ T_Q/\eta_{u,l} \cdot \{ \ 1 + \eta_{u,l} \cdot n_B(\nu_{u,l}) + 1/g_M \cdot [2 \cdot e_0 \cdot <i_M> \cdot R_M/(k_B \cdot T_Q) + T_{IF}/T_Q] \ \} \qquad (12)$$

The important finding is that the system temperature does not depend on the bandwidth. This together with the linear behavior of the effective noise temperature with respect to additional external sources makes it a most useful definition for heterodyne systems. One particular difference to the direct detection equations Eq.(4) or (5) is that there is a minimum noise contribution - the “quantum limit” - according to the first spontaneous emission term in the brackets in Eq.(12), which persists even at zero contribution of all other noise terms.[8]

The influence of the IF-amplifier noise contribution we may now try to estimate. At 500 GHz for example the quantum limit T_Q is 24 K, while a typical noise temperature of a broadband cooled HEMT IF-amplifier is in the range of 5 K. According to Eq.(12) the amplifier contribution is given by

$$1/g_M \cdot T_{IF}/T_Q = 1/0.1 \cdot 5/24 = 2.08 \ (g_M = \text{-10 dB}).$$

It adds therefore more than 200% of the spontaneous emission term to the final noise temperature, which is by no means negligible. This changes when going into the mid-infrared. The quantum limit at 10 µm is 1440 K, so that a HEMT-amplifier with 5 K noise temperature and similar mixer gain yields a contribution to the system noise of about 3.5 % of the quantum limit (and becomes much less with very efficient mixer detectors). Even a 100 K amplifier is still acceptable. Therefore, the IF-amplifier has much less influence on the system noise temperature at very high frequencies as long as the mixer gain is not too small.

For a dual sideband (DSB) mixer we find:

$$T_{Sys}(DSB) \ = \ T_Q/(\eta_u+\eta_l) \cdot \{ \ 1 + 2 \cdot \eta_H^{eff} \cdot n_B(\nu_{LO}) + 1/g_M \cdot [2 \cdot e_0 \cdot <i_M> \cdot R_M/(k_B \cdot T_Q) + T_{IF}/T_Q] \ \} \qquad (13)$$

Here we have used the fact that the two sideband frequencies are nearly identical with the LO-frequency: $\nu_{u,l} = \nu_{LO} \pm \nu_{IF} \approx \nu_{LO}$. η_H^{eff} is the average sideband efficiency with $\eta_H^{eff} = ½ \cdot (\eta_u+\eta_l)$. Note that the background contribution comes now from both sidebands, but only from one spatial mode in each sideband. Therefore, a factor of 2 appears in the background term. In case one has a line signal in only one of the sidebands while observing with a dual sideband mixer, one has to recalibrate and we can define the single sideband system temperature with:

$$T_{Sys}(SSB_{u,l}) \ = \ (\eta_u+\eta_l)/\eta_{u,l} \cdot T_{Sys}(DSB) \ =$$

[8] Actually, if one inserts an amplifier in front of a direct detector, this quantum limit appears as well due to the spontaneous emission in such amplifier (see e.g. [17], [18]).

$$= T_Q/\eta_{u,l} \cdot \{ 1 + 2\cdot\eta_H^{eff}\cdot n_B(\nu_{LO}) + 1/g_M \cdot [2\cdot e_0\cdot<i_M>\cdot R_M/(k_B\cdot T_Q)+T_{IF}/T_Q] \}$$

This single sideband noise temperature is not identical with Eq.(12), since the background contribution is not the same. The improvement in sensitivity of a single sideband mixer depends solely on the value of the background mode occupation number n_B, which may be large at low frequencies and is negligible in the infrared. Whether or not one should develop a single sideband mixer depends only on the amount of background radiation seen with the receiver. Despite the problem of sideband de-convolution, when observing dense molecular spectra, it is not worthwhile to build a single sideband mixer in the mid-infrared, since the background mode occupation numbers are typically too small to be of significance.

When using the definition of the *NEP* from earlier we can now introduce the single sideband *NEP* for a dual sideband heterodyne system as well [9]:

$NEP(SSB_{u,l}) = \surd 2 \cdot k_B\cdot T_{Sys}(SSB_{u,l}) \cdot \surd(\delta_{Res}/q)$ and $NEP(DSB) = 2^{3/2} \cdot k_B\cdot T_{Sys}(DSB_{u,l}) \cdot \surd(\delta_{Res}/q)$

In the ideal case of 100% quantum efficiency and without any further noise contributions from amplifier, shot noise, or background we have (while assuming $q = 1$, which is equivalent to an ideal filter):

$$NEP(SSB_{u,l})_{Min} = NEP(DSB)_{Min} = \surd 2\cdot k_B\cdot T_Q \cdot \surd\delta_{Res}$$

This defines the minimum *NEP* achievable with a heterodyne system. For example we find at 500 GHz: $NEP_{Min} = 1.9\cdot 10^{-19}$, and at 30 THz: $NEP_{Min} = 8.9\cdot 10^{-17}$ Watt/Hz$^{1/2}$ while assuming a frequency resolution of $R = 3\cdot 10^6$ (δ_{Res} = 10 MHz @ 30 THz, or δ_{Res} = 167 kHz @ 500 GHz).

2.a. The shot noise contribution

As was mentioned before, the constant *c* in Eq.(8) is dependent on the type of mixer detector and is usually unknown. But we can make some estimate for a classical square-law mixer. Starting with the simplest classical treatment of the heterodyne process with a single frequency signal component, we can write for the source current i_M of a square law mixer detector:

$$i_M = \eta_H \cdot e_0/h\nu_{LO} \cdot \{P_{LO} + P_S + 2 \cdot \surd[P_{LO}\cdot P_S]\cdot cos(2\pi\cdot\nu_{IF}\cdot t)\} + i_B$$

i_B is the bias current, which is needed at the best operating point of the mixer. η_H is the quantum efficiency of the mixer, and P_S and P_{LO} are the signal and the LO power respectively. The signal frequency is again assumed to be nearly identical with the LO frequency. The equation describes the ideal heterodyne process, which assumes that there are no additional losses of IF-signal at frequency ν_{IF}, which may occur due to time constants or impedance mismatch for example. The maximum photo current is given for $\eta_H = 1$, and we find the maximum photo-responsivity $r_0 = e_0/h\nu_{LO} \approx 485$ Amp/Watt at 500 GHz and $r_0 = e_0/h\nu_{LO} \approx 8.1$ Amp/Watt at 10 µm wavelength. Very good mixers have a quantum efficiency near or above 50%, so that realistic responsivities $r = \eta_H \cdot e_0/h\nu_{LO}$ are about half of these values.

The mean generated IF-power is then equal to:

$$P_{IF} = <i_{AC}^2>\cdot R_M = 4 \cdot \eta_H^2 \cdot (e_0/h\nu_{LO})^2 \cdot P_{LO} \cdot P_S \cdot R_M \cdot <cos^2(2\pi\cdot\nu_{IF}\cdot t)> =$$
$$= 2 \cdot \eta_H^2 \cdot (e_0/h\nu_{LO})^2 \cdot P_S \cdot P_{LO} \cdot R_M$$

with R_M the dynamic (real valued) impedance of the mixer. On the other hand, we can write for the mixer gain, which is the ratio of IF-power and absorbed input signal power:

$$g_M = P_{IF}/(\eta_H \cdot P_S) = 2 \cdot \eta_H \cdot (e_0/h\nu_{LO})^2 \cdot P_{LO} \cdot R_M$$

With an impedance $R_M = 50\ \Omega$ and a photo-current of about 2 mA ($P_{LO} \approx 0.5$ mWatt, $\eta_H = 0.5$) the mixer gain is larger than 1 at 10 µm wavelength. The same we find at 500 GHz with an LO power

[9] Frequently, the sensitivity of a heterodyne system is described by the spectral density of the noise power of the system (in Watt/Hz), which is identical with the system noise temperature T_{Sys} times the Boltzmann constant k_B, while it is also called "*NEP*". This should not be confused with the real noise equivalent power *NEP*, as is unfortunately sometimes done in the literature!

near 140 nWatt ($<i_{LO}> \approx 34$ µA) [10]. For the ratio of the IF-photon number n_{IF} to the absorbed signal photon number n_S, which is the photon gain G_M of the mixer itself, we have:

$$G_M = \nu_{LO}/\nu_{IF} \cdot g_M,$$

and we find therefore for the mixer photon gain at 500 GHz or 30 THz respectively and 1 GHz IF-frequency:

$$G_M \approx 800 \text{ (500 GHz) and } \approx 5\cdot10^4 \text{ (30 THz)}$$

In the mid-IR there is enormously high photon gain, and it differs drastically from the values seen with mm/submm mixers.

In order to estimate what the expected contribution to the noise temperature should be, we have now (see Eq.(13))

$$1/g_M \cdot 2\cdot e_0\cdot<i_M>\cdot R_M/(k_B\cdot T_Q) = 2\cdot e_0/h\nu_{LO} \cdot (c\cdot<i_{LO}>+<i_B>)\cdot R_M / [2\cdot r_0\cdot<i_{LO}>\cdot R_M] =$$
$$= c + <i_B>/<i_{LO}>,$$

while using $<i_M> = c\cdot<i_{LO}>+<i_B>$ (see Eq.(6).)

Note here that the shot noise contribution does not depend on the mixer impedance R_M or the coupling efficiency to the IF-amplifier, since both, the shot noise and the signal are affected identically by eventual coupling losses. Consequently, the details of the mixer characteristics do not play a significant role here. In total, we have now a noise temperature of

$$T_{Sys}(SSB)_{u,l} = T_Q/\eta_{u,l} \cdot \{ 1 + c + \eta_{u,l}\cdot n_B(\nu_{LO}) + <i_B>/<i_{LO}> + 1/g_M \cdot T_A/T_Q \}$$

It is also now evident that for $c = 1$ the minimum SSB noise temperature ($\eta_{u,l} = 1$) is exactly identical with $2\cdot T_Q$, since the smallest possible shot noise term is equal to unity ($<i_B> = 0$, $T_A = 0$). In some cases, the quantum limit of IR-heterodyne systems is derived in literature from the shot noise term only, while the spontaneous emission term is not considered (see e.g. [2], [3], or [4]). To our understanding this is not a proper description of the physics of mixing, because it could lead to below quantum limit performance of SIS mixers for example.

Similar as before, for a double sideband mixer we find accordingly:

$$T_{Sys}(DSB) = T_Q/(\eta_u+\eta_l) \cdot \{ 1 + c + 2 \cdot \eta_H^{eff}\cdot n_B(\nu_{LO}) + <i_B>/<i_{LO}> + 1/g_M \cdot T_A/T_Q \} \quad (14)$$

Therefore, for a "classical" mixer the minimum DSB noise temperature is equal to T_Q and not equal to $T_Q/2$, as is frequently stated. But, at mm- and submm-frequencies SIS-mixers are the best choice, and for those we have c = 0 and find a theoretical limit of $T_Q/2$.[11] Certainly, this is never observed, since the influence of background radiation and IF-amplifier is usually dominating the noise budget.

We can now estimate the expected best noise temperatures of a realistic 500 GHz SIS receiver ($\eta_H = 0.5$, $c = 0$, $n_B(\nu) = f/[e^{T_Q/T} - 1] = 1.18$, $<i_B>/<i_{LO}> \approx 1$, $T_A = 5$ K, $T_Q = 24$ K) with

$$T_{Sys}(DSB) = 24/(2\cdot0.5) \cdot \{ 1 + 0 + 2\cdot0.5\cdot1.18 + 1 + 5/24 \} \approx 80 \text{ K}$$

This fits fairly well to the best observed results of SIS mixers in this frequency range. One can state that the background at 295 K and the bias current contribute equally to the noise temperature. Certainly, the estimate of the quantum efficiency at $\eta_H = 0.5$ is an educated guess, but it includes the losses in front of the mixer, so that it is probably not too far from reality.

In the mid-infrared heterodyne systems like THIS or HIPWAC should have:

$$T_{Sys}(DSB) = 1440/(2\cdot0.5) \cdot \{ 1 + 1 + 2\cdot7.8\cdot10^{-4} + 0 + 100/1440 \} \approx 2980 \text{ K}$$

[10] As is shown by Tucker and Feldman [19], the gain of a classical mixer should not exceed a value of 1. But our very simple model ($i_M \sim u^2$, u the applied voltage, negative and positive!) does not describe a real mixer. We therefore find a mixer gain which can in principle become infinite. On the other hand, since the quantum energy is fairly high at 10 µm wavelength, the classical limit does not apply anyway. It is also shown in [19] that a quantum mixer might exhibit appreciable gain under certain circumstances.

[11] At this place it is important to understand that the quantum limit is the result of noise seen at the IF-output of the mixer. It is a bit confusing, when sharing it between the two sidebands of the mixer. It is certainly impossible that only half a noise photon in the IF comes from each sideband!

with $\eta_H \approx 0.5$, $c = 1$, $n_B(\nu) = f/[e^{T_Q/T} - 1] = 7.8\cdot10^{-4}$, $<i_B>/<i_{LO}> \approx 0$, $T_A \approx 100$ K. It matches well to observed DSB noise temperatures near or below 3000 K at 10 µm ([22]).[12] It is also evident that the amplifier contribution is practically insignificant, which is very different for low frequency mixers. By the way, there are good arguments that for a photo diode the value of *c* may be smaller than unity [21], so that a lower noise temperature might exist, although it should not be very different for the typical HgCdTe-mixers in use in the 10 µm wavelength region. It therefore seems rather unlikely to reach values below 2000 K unless there is significant improvement of the quantum efficiency η_H, which also depends on the efficiency of the optics in front of the mixer. It is important to note that at high values of LO current the additional bias current $<i_B>$ becomes negligible in comparison with the photo-current $<i_{LO}>$, so that any increase of the LO power does not improve the noise temperature anymore.

One can now also understand why the system noise temperature rapidly deteriorates with low LO power. Usually, the overall current through the mixer needs to be kept at nearly constant level in order to remain at the best operating point on the I/V-curve of the mixer. But the ratio of bias current $<i_B>$ and LO photo-current becomes then fairly large. For example, at 10 µm and with an LO power of, say, 50 µWatt, the LO current is approximately 0.2 mA (see above), while the bias current must add 1.8 mA for a total of 2 mA detector current again. From this we find a noise temperature of

$$T_{Sys}(DSB) \approx T_Q/(2\cdot0.5)\cdot\{\,1 + c + <i_B>/<i_{LO}>\,\} = ½\cdot1440/0.5\cdot\{\,1 + 1 + 1.8/0.2\,\} \approx 15800\text{ K}$$

This is rather close to the results found when pumping our HgCdTe mixer with insufficient power from Lead-salt tunable diode lasers for example (see e.g. [23]). This is a good explanation why attempts to pump IR-mixers with low power Lead-salt lasers have not yet been very successful.

3. Concluding Remarks

The discussion above suggests that the sensitivity of heterodyne detection is competitive with direct detection methods at high frequency resolution even in the mid-infrared. Since the experimental methods for the determination of sensitivities of heterodyne and direct detection systems are not exactly comparable, it is still difficult to find clear answers when considering different instruments, and, to our knowledge, a simultaneous comparison of such instruments during real observations has not yet been performed. There is also the argument of imaging capability of direct detection instruments, since spectrometers like TEXES [15] or FIFI [28] are equipped with many pixel array-detectors, although the number of usable image pixels is rather limited, particularly at high frequency resolution. On the other hand, multi-pixel imaging has also been established for heterodyne instruments (see e.g. [26], [27] or others) so that the multiplex-advantage may not differ very much between both methods. Certainly, due to the higher frequency resolution the total frequency coverage of a heterodyne instrument is smaller, but the total number of available resolution elements is quite comparable. No doubt, also mid-IR heterodyne receivers can be developed with several mixer detectors all pumped with one laser.

The discussion here may be useful for applications in interferometry in the mid-infrared. In case, high frequency resolution needs to be combined with high spatial resolution, it is clear that the heterodyne method has the additional advantage that the IF signal can be amplified to any desired level, so that losses are avoided when distributing the received signal power into many correlators of a large number of baselines in a multi-telescope arrangement. This gives the heterodyne method an additional advantage (see e.g. [11]). Usually, when observing at short wavelengths, the correlation of signals between pairs of telescopes is sampled with a direct detection system.

[12] Recently, an uncorrected system temperature of THIS has been measured with 2400 K, which is probably due to an improved quantum efficiency of our mixer detector. Nevertheless, with this result we are already approaching the theoretical limit when considering all optical losses in the spectrometer.

Therefore, to supply each baseline, the number of photons collected with each of the N telescopes must be divided up so that from the initial n_S signal photons per telescope only n_S/N photons are left for the extraction of the interference signals. For example, if the telescopes have an aperture of *A* m^2, the effective collecting area for the interferometry is only *A/N* per baseline. In addition, there are also losses in the delay-lines, which may amount to more than 50%. This reduces the effective collecting area by at least another factor of two. For the Very Large Telescope (VLT) in Chile it means that the effective area per telescope is less than 6.6 m^2 instead of about 50 m^2! This corresponds to a single dish diameter of less than 3 m. In comparison, there is a factor of more than 8 to the advantage of a heterodyne interferometer, which fully compensates for an eventually initial loss due to the noise temperature of heterodyne instruments. With larger number of telescopes this advantage increases further. Certainly, the argument applies only when considering high spectral resolution and rather small bandwidth at the same time. At low frequency resolution the situation is certainly different.

3.a. Calibration

Obviously, there are differences in the interpretation of sensitivities between the heterodyne and the direct detection world. The usage of noise temperatures for heterodyne systems is based on the assumption that only the wave fluctuation noise contributes, whereas the shot noise is assumed to dominate in direct detectors. This means that the methods for measuring the sensitivity of the instruments must be different. Heterodyne systems are typically characterized by the so-called "y-factor method", which assumes that the radiometer equation is strictly valid. It requires that the standard deviation of the noise fluctuations is proportional to the mean output of the instrument. Therefore, a measurement of the generated system noise power also provides information about the sensitivity. The method is based on two measurements of thermal loads, one "hot" and one "cold" at brightness temperatures *J(Hot)* and *J(Cold)* [13] with observed output signals *S(Hot)* and *S(Cold)*. From the ratio of the two signals the noise contribution of the system is now calculated with:

$$T_{Sys}(DSB) \quad = \quad \frac{J(Hot) - y \cdot J(Cold)}{y-1} \qquad y = \frac{S(Hot)}{S(Cold)}$$

Once again, the derived system noise temperature is a measure for the spectral power density of the receiver noise and has nothing to do with a real temperature. The derived predictions for instruments like THIS and HIPWAC are surprisingly close to the experimental data. The question, whether the shot noise can be made smaller than found with the naive assumptions used above, is still to be investigated. It may be possible that double sideband noise temperatures can be reached, which are significantly below the present best values.

The *y*-factor method is not applicable for direct detectors, because the generated noise power is not proportional to the average detector output, as is required for the two signals *S(Hot)* and *S(Cold)*. Therefore, in order to obtain comparable results, another experimental method for the characterization of direct detector systems must be applied, which is leading to the same information as the *y*-factor method used for heterodyne receivers. One method could be, to do this in two steps:

1. Determine the response of the detector output to known inputs (calibration load at known brightness temperatures)

[13] It is frequently proposed to use the Callan-Welton formula [24] instead of Eq.(6) for the calculation of the brightness temperatures *J* of the two loads [25]:

$$J = T_Q \cdot [1//(e^{T_Q/T} - 1) + ½]$$

In effect, this reduces the calculated DSB system noise temperature by exactly $T_Q/2$, and it appears like an improvement of the system. But, when dealing with such system temperature one assumes that the mixer could exist while all vacuum fluctuations are turned off, since the spontaneous emission term in Eq.(7) is the consequence of their influence. It is rather unusual to neglect the effects of the vacuum interaction for the description of a quantum mechanical system, and we therefore propose that this approach should not be used.

2. Determine the standard deviation of the fluctuations of the system output at zero input (cold calibration load).

When calibrating the standard deviation of the noise fluctuations, the sensitivity of the system is calculated. It leads to figures which are fully equivalent to the system noise temperatures as they are provided for heterodyne systems and the conversion to *NEP*s or *NEFD*s or else is simple. By the way, the same procedure could also be used for heterodyne systems.

3.b. Signal to Noise Ratio: a Comparison

For a better understanding of advantages or disadvantages of the two methods, the formulas derived before may be used for a comparison between heterodyne and direct detection sensitivities. But some realism must be applied in order to obtain credible results. In both cases we assume signals from point-like sources. For direct detection systems it is important to take into consideration that the detection efficiency at high frequency resolution is usually fairly small in comparison with heterodyne systems. In addition, it is also essential that the frequency resolution can not be made arbitrarily high so that frequency dilution effects become important when observing narrow line signals. On the other hand, heterodyne receivers suffer strongly from the quantum limit when observing at very high frequencies, but the resolution can be optimized on the other hand. As an example, we made the attempt to include all these parameters into one graph in Fig.3, but, in order to include the frequency dilution problem it is now more informative to consider the signal to noise ratio (*SNR*).

For simplicity the spectrometer response function and the signal line shape are assumed as Gaussians. Thus the frequency dilution can be described by:

$$S \ = \ \frac{S_0}{\sqrt{1+(\Delta\nu/\delta\nu)^2}} \ \sim \ \frac{1}{\sqrt{1+(\Delta\nu/\delta\nu)^2}}$$

Δν is the halfwidth (FWHM) of the spectrometer filter function and *δν* the corresponding signal width. *S₀* is the true amplitude of the signal and S the reduced amplitude observed according to the frequency dilution.

For the noise we have in both cases the radiometer equation with:

$$\Delta T \ = \ \frac{T_{Sys}}{\sqrt{B_{Fl}\cdot t_{Int}}} \ \sim \ \frac{T_{Sys}}{\sqrt{\Delta\nu}}$$

The *SNR* is therefore in case of direct detection at fixed resolution *Δν*:

$$SNR(Dir) \ \sim \ \frac{1}{T_{Sys}^{Dir}}\cdot\frac{\sqrt{\Delta\nu}}{\sqrt{1+(\Delta\nu/\delta\nu)^2}}$$

In the heterodyne case one can optimize the *SNR* by adjusting the resolution width *Δν*, which leads to *Δν* = *δν*, and we have:

$$SNR(Het) \ \sim \ \frac{1}{T_{Sys}^{Het}}\cdot\sqrt{\frac{\Delta\nu}{2}}$$

For the graphs in Fig.3 we have used Eqs.(5) and (14) for an estimate of the noise temperatures while assuming that the systems are only background limited and other noise contributions are zero. As a starting point, the system temperatures as are found with TEXES and THIS at 10 µm wavelength are used and extrapolated to the other wavelengths using Eq.(3) for the background contributions.

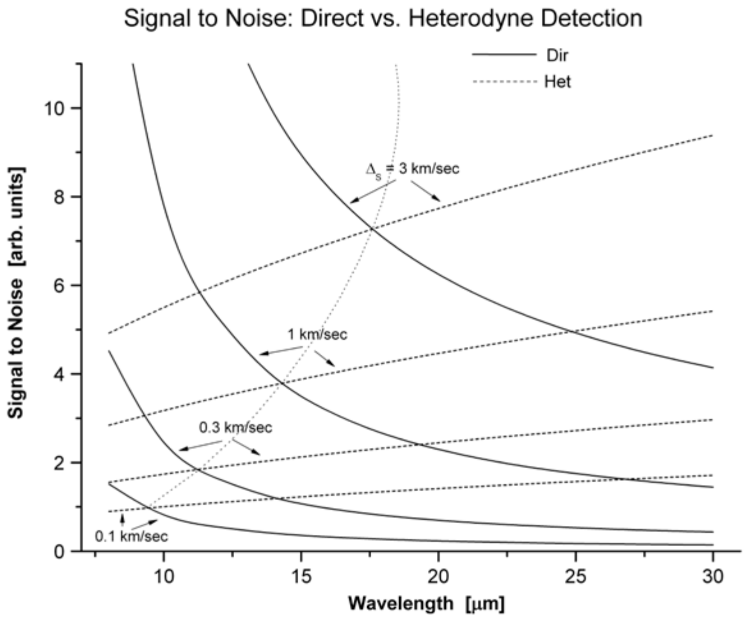

Fig.:3

Signal to noise ratio of direct detection (solid lines) and heterodyne detection (dashed lines) vs. wavelength for different linewidth of a signal. For direct detection the signal to noise is extrapolated to 30 µm using a constant frequency resolution of 300 MHz at all wavelengths (R = 10^5 at 10 µm). The signals therefore become frequency diluted at small signal width. For heterodyne detection the resolution is optimized to equal the frequency width of the signal. In both cases the influence of thermal background radiation is included. The dotted line connects the crossing points of both systems. To the left, direct detection is more sensitive, to the right heterodyne detection is advantageous.

The *SNR* is shown for various widths of the detected signal. It is interesting to note that the heterodyne *SNR* increases with increasing wavelength, mainly because the quantum limit decreases, while the direct detection *SNR* decreases, since the background contribution increases. One might expect that a heterodyne receiver is always less sensitive due to the influence of the quantum limit, because it sees similar background. But this is largely compensated by the very different coupling efficiencies η. Therefore, there exists a distinct crossover point for a given signal width. The dotted line in the plot connects these points as a function of signal width. If the background exceeds the quantum limit, heterodyne reception is advantageous, because of the much higher coupling efficiency. This is valid for the longer wavelength regime. At shorter wavelength, the heterodyne system suffers from the increased value of the quantum limit so that direct detection is more sensitive. In addition, the background contribution shrinks drastically at shorter wavelength. Above about 18.5 µm the system noise temperature of the heterodyne system itself becomes smaller than that of the direct detection system, i.e. the heterodyne reception is more sensitive at any reasonable resolution. This certainly depends strongly on the coupling efficiency η of direct detection spectrometers and might shift to longer wavelength if improved. When operating with a cooled telescope in space, the situation also changes, and it becomes rather likely that the detector parameters will decide about the sensitivity at the end.

It should be noted that the phrase "line width" stands for the width of the narrowest features in the spectrum one wants to observe. For example, absorptions from the cold gas of interstellar clouds may be responsible for narrow features, which are visible in much broader emission lines from star forming regions. It is therefore quite common that a resolution in the range of 1 km/sec or below is necessary for such observations. Similar, signals from high altitude in planetary atmospheres require usually even higher resolution. As is visible in Fig.3, the sensitivity of heterodyne detection exceeds always that of direct detection in the mid-IR, if very narrow features below 0.1 km/sec (i.e. 10 MHz resolution bandwidth at 10 µm) are observed. This is for example important for wind measurements on solar planets, as are reported in [8], [9] or [22].

References:

[1] I. Harris, “Coherent and incoherent detection at submillimeter and far-infrared wavelengths,” in Coherent Detection at Millimeter Wavelengths and their Applications, P. Encrenaz, C. Laurent, S. Gulkis, E. Kollberg, and G. Winnewisser, eds., Les Houches Series (Nova Science, New York, 1991), pp. 7–34

[2] R.H. Kingston; "Detection of Optical and Infrared Radiation", Springer Series in Optical Sciences, D. L. MacAdam ed., (1978)

[3] T.G. Blaney, “Signal-to-noise ratio and other characteristics of heterodyne radiation receivers,” Space Sci. Rev. 17, 691–702 (1975)

[4] B. Parvitte, V. Zéninari, C. Thiébeaux, A. Delahaigue, D. Courtois; "Infrared laser heterodyne systems", Spectrochimica Acta A 60, 1193-1213 (2004)

[5] Th. Kostiuk; "Heterodyne Spectroscopy in the Thermal Infrared Region: A Window on Physics and Chemistry", in proceedings of NASA's "International Thermal Detectors Workshop (TDW 2003)", NASA CP-2004-212748; Rept-2004-00980-0, page 7-1 to 7-7 (2004)

[6] D.D.S. Hale, M. Bester, W.C. Danchi, W. Fitelson, S. Hoss, E.A. Lipman, J.D. Monnier, P.G. Tuthill, C.H. Townes; "The Berkeley Infrared Spatial Interferometer: A Heterodyne Stellar Interferometer for the Mid-Infrared", ApJ 537, 998-1012 (2000)

[7] G. Sonnabend, M. Sornig, P. Krötz, D. Stupar, R. Schieder; "Ultrahigh Spectral Resolution Observations using the Cologne Tuneable Heterodyne Infrared Spectrometer", submitted to JQSRT (2007)

[8] T. Kostiuk, T.A. Livengood, T. Hewagama, G. Sonnabend, K.E. Fast, K. Murakawa, A.T. Tokunaga, J. Annen, D. Buhl, F. Schmülling; "Titan's stratospheric zonal wind, temperature, and ethane abundance a year prior to Huygens insertion", Geophys. Res. Lett, 32 no.22, L22205 (2005)

[9] G. Sonnabend, K. Fast, M. Sornig, P. Kroetz, R. Schieder; "High spatial resolution mapping of Mars mesospheric zonal winds by infrared heterodyne spectroscopy of CO_2", Geophysical Research Letters, VOL. 33, L18201 (2006)

[10] P. Felgett, R.C. Jones, R.Q. Twiss; "Fluctuations in photon streams", Nature 184, 967-969 (1959)

[11] J. Zmuidzinas; "Cramér-Rao sensitivity limits for astronomical instruments: implications for interferometer design", J.Opt.Soc.Am. A, Vol.20 No.2, 218-233 (2003)

[12] J. Zmuidzinas; "Thermal noise and correlations in photon detection", Appl. Optics 42, 25, 4989-5008 (2003)

[13] R.H. Dicke; "The measurement of thermal radiation at microwave frequencies", Rev. Sci. Instrum. 17, 268-275 (1946)

[14] R. Loudon; "The quantum theory of light", Oxford University Press, pp. 210-225 (1973)

[15] J.H. Lacy, M.J. Richter, T.K. Greathouse, D.T. Jaffe, Q. Zhu; "TEXES: A Sensitive High-Resolution Grating Spectrograph for the Mid-Infrared", PASP 114, 153-168 (2002)

[16] J.Zmuidzinas; "The Role of Coherent Detection", Proceedings of the NASA Second Workshop on New Concepts for Far-Infrared and Submillimeter Space Astronomy, College Park/MD, D. Benford and D. T. Leisawitz, eds., document CP-2003-212233 (NASA, 2003), pp. 329-341 (2002)

[17] H. Heffner; "The fundamental noise limit of linear amplifiers", Proc. IRE 50, 1604-1608 (1962)

[18] C.M. Caves; "Quantum limits on noise in linear amplifiers", Phys. Rev. D 26-8, 1817-1839 (1982)

[19] J.R. Tucker, M.J. Feldman; "Quantum detection at millimetre wavelengths", Rev. Mod. Phys. 57,4, 1055-1113 (1985)

[20] J.D. Kraus; "Radio Astronomy", 2nd edition, Cygnus-Quasar Books, pp. 7-25 – 7-27 (1986)

[21] A. Uhlir, Jr.; "Shot Noise in p-n Junction Frequency Converters", Bell Syst. Tech. J. 37, 951-988 (1958)

[22] G. Sonnabend, D. Wirtz, V. Vetterle, R. Schieder; "High-resolution observations of Martian non-thermal CO_2 emission near 10 µm with a new tuneable heterodyne receiver", Astronomy and Astrophysics, 435, 1181 (2005)

[23] D.Wirtz, G. Sonnabend, R. Schieder; "THIS: a tuneable heterodyne infrared spectrometer", Spctrochimica Acta A 58, 2457-2463 (2002)

[24] H.B. Callen and T.A. Welton, "Irreversibility and generalized noise," Phys. Rev., vol. 83 no.1, 34-40 (1951)

[25] A. R. Kerr, M. J. Feldman, and S. K. Pan, “Receiver noise temperature, the quantum noise limit, and the role of the zero-point fluctuations,” in Proceedings of the Eighth International Symposium on Space Terahertz Technology (1997), pp. 101–111. Originally printed as Electronics Division Internal Report No. 304 (National Radio Astronomy Observatory, Charlottesville, Va. 22903), (1996)

[26] U.U.Graf, S. Heyminck, E.A.Michael, S. Stanko, C.E. Honingh, K. Jacobs, R.T. Schieder, J. Stutzki, B. Vowinkel; "SMART: The KOSMA Sub-Millimeter Array Receiver for Two frequencies", Millimeter and Submillimeter Detectors for Astronomy. Edited by Phillips, Thomas G.; Zmuidzinas, Jonas. Proceedings of the SPIE, Volume 4855, pp. 322-329 (2003)

[27] K.-F. Schuster, C. Boucher, W. Brunswig, M. Carter, J.-Y. Chenu, B. Foullieux, A. Greve1, D. John, B. Lazareff, S. Navarro, A. Perrigouard, J.-L. Pollet, A. Sievers, C. Thum, H. Wiesemeyer; "A 230 GHz heterodyne receiver array for the IRAM 30 m telescope", A&A 423, 1171-1177 (2004)

[28] A. Poglitsch, J.W. Beeman, N. Geis, R. Genzel, M. Haggerty, E.E. Haller, J. Jackson, M. Rumitz, G.J. Stacey, C. H. Townes; "The MPE/UCB far-infrared imaging Fabry-Perot interferometer (FIFI)", Int. J. IR and MM Waves 12, no.8, 859-884 (1991)