Millimeter Interferometry

Proceedings of the

"IRAM Millimeter Interferometry Summer School"

edited by S.Guilloteau

Institut de Radio Astronomie Millimétrique
300 rue de la piscine
F-38406 Saint Martin d'Hères Cedex, FRANCE

In Memoriam

This book is dedicated to the memory of our esteemed colleagues

- Bernard AUBEUF
- Francis GILLET
- Henri GONTARD
- Roland PRAYER
- Patrick VIBERT

and of the other 15 persons who died on 1st of July, 1999, in the tragic accident of the cable car giving access to the IRAM interferometer site on Plateau de Bure.

Bernard AUBEUF and Patrick VIBERT had been with the IRAM interferometer since the very beginning of its construction. With the help of Henri GONTARD, they had assumed the difficult task of handling the cable car, the general maintenance of the Plateau de Bure buildings, as well as long days of snow clearing to allow the configuration changes in winter. Francis GILLET and Roland PRAYER had been regularly joining the IRAM staff for the summer maintenance, and for the construction of the antennas.

15 employees of companies working as contractors for IRAM (either on the baseline extension, or on facilities like telephone equipment) also lost their lives in that accident.

Without their efforts, millimeter astronomy would never have come to the Plateau de Bure site. We wish to dedicate these proceedings to their memory.

Preface

This book contains the proceedings from a summer school held at IRAM in Grenoble from September 14 to September 18, 1998. The school was attended by more than 50 participants from abroad. Not all lectures have been put in written form: the school also included a general introduction to millimeter astronomy by Dr. Clemens THUM (IRAM). A visit of the IRAM interferometer on Plateau de Bure was an essential part of the school, which gave an opportunity to the attendants to confront the concepts described in the lectures to their practical implementation on the site.

The school would not have occurred without the dedication of Mrs Catherine BERJAUD, who kindly took care of all logistics problems before and during the school.

Special thanks are also due to the staff of the Plateau de Bure interferometer, who made possible a visit to the site despite the inherent difficulties to handle such a large group of persons.

We also thank Alain PERRIGOUARD and Roger AHTCHOU for help with the video equipment which gave us a lot of fun during the "real time" sessions.

Contents

In Memoriam **iii**

Preface **v**

1 Radio Antennas **1**
1.1 Introduction . . . 1
1.2 Basic Principles . . . 2
1.3 The perfect Single–Dish antenna . . . 3
1.4 The real Single–Dish Antenna . . . 7
1.4.1 Sytematic Deformations: Defocus, Coma, Astigmatism . . . 7
1.4.2 Random Errors . . . 10
1.5 Radiometric Relations . . . 13

2 The interferometer principles **15**
2.1 Basic principle . . . 15
2.2 The Heterodyne Interferometer . . . 17
2.2.1 Source Size Effects . . . 18
2.2.2 Finite Bandwidth . . . 19
2.3 Delay Tracking and Frequency Conversion . . . 19
2.4 Fringe Stopping and Complex Correlator . . . 20
2.5 Fourier Transform and Related Approximations . . . 21
2.6 Array Geometry & Baseline Measurements . . . 22

3 Receivers : an overview for non-specialists **25**
3.1 Introduction . . . 25
3.2 Coupling optics . . . 25
3.3 Why we need heterodyne receivers . . . 27
3.4 Local oscillator system . . . 27
3.5 Local oscillator injection . . . 27
3.6 Photon-assisted tunneling . . . 28
3.7 Mixer . . . 29
3.8 Cryostat . . . 30
3.9 Actual receivers . . . 31

4 Cross Correlators **35**
4.1 Introduction . . . 35
4.2 Basic Theory . . . 37
4.3 The Correlator in Practice . . . 38
4.3.1 Digitization of the input signal and clipping correction . . . 39
4.3.2 Time lag windows and spectral resolution . . . 43
4.3.3 Main limitations . . . 43
4.4 The correlator on Plateau de Bure . . . 47
4.4.1 The current system . . . 48

4.4.2 Future improvements . . . 48
4.5 Appendix . . . 49
4.5.1 Summary of definitions . . . 49
4.5.2 Clipping correction for 4-level quantization . . . 49

5 LO System and Signal Transport 51
5.1 An Heterodyne Interferometer . . . 51
5.1.1 The simple interferometer . . . 51
5.1.2 The heterodyne interferometer . . . 51
5.1.3 Frequency conversion . . . 51
5.1.4 Signal phase . . . 52
5.2 Delay lines requirements . . . 54
5.2.1 Single side band processing in a finite bandwidth . . . 54
5.2.2 Double side band system . . . 55
5.3 Side band separation . . . 55
5.3.1 Fringe rate method . . . 55
5.3.2 Phase switching method . . . 55
5.4 The PdB Signal and LO transport system . . . 56
5.4.1 Signal path . . . 56
5.4.2 LO generation . . . 56
5.4.3 Further signal processing . . . 58
5.4.4 Phase stability requirements . . . 58
5.4.5 Cable electrical length control . . . 59
5.5 Next generation instruments . . . 59

6 The Plateau de Bure Interferometer 61
6.1 History . . . 61
6.2 Description . . . 61
6.3 Array operation . . . 63
6.3.1 Array calibration . . . 63
6.3.2 Array observations . . . 66
6.4 Proposal submission and contact people . . . 68

7 Band pass and Phase Calibration 71
7.1 Definitions and formalism . . . 71
7.1.1 Baseline based vs antenna based gains . . . 71
7.1.2 Gain corrections . . . 72
7.2 Bandpass calibration . . . 73
7.2.1 Bandpass measurement . . . 73
7.2.2 IF passband calibration . . . 73
7.2.3 RF bandpass calibration . . . 74
7.2.4 Side band calibration . . . 75
7.3 Phase calibration . . . 75
7.3.1 Phase referencing by a nearby point source . . . 76
7.3.2 Phase referencing by a point source in the primary beam . . . 77
7.3.3 Phase referencing using another band or another frequency . . . 77

8 Atmospheric Absorption 79
8.1 The Atmosphere . . . 79
8.1.1 Constituents of the atmosphere . . . 79
8.1.2 Thermodynamics of the air . . . 80
8.1.3 Hydrostatic equilibrium . . . 81
8.1.4 Water . . . 83
8.1.5 Photodissociation products . . . 83

8.2 Spectroscopy of H_2O, O_2, O_3 . . . 83
8.2.1 Water vapor . . . 83
8.2.2 Molecular oxygen . . . 84
8.2.3 Ozone . . . 85
8.3 Propagation of a wave in the atmosphere – Line shapes . . . 85
8.4 The atmospheric absorption spectrum at millimeter wavelengths, ATM . . . 87
8.5 Correction for atmospheric absorption, T_A^* . . . 87
8.5.1 Simplest case . . . 89
8.5.2 More realistic case . . . 90
8.5.3 General case . . . 90

9 Atmospheric Fluctuations **91**
9.1 Introduction . . . 91
9.2 Hydrodynamical basics of turbulent motion . . . 93
9.3 Statistical properties of turbulence . . . 96
9.4 Remote sounding techniques . . . 98
9.5 Current phase correction at IRAM . . . 99
9.6 Phase correction during off-line data reduction . . . 102
9.7 Frequently asked questions . . . 103

10 Amplitude and Flux Calibration **105**
10.1 Definition and Formalism . . . 105
10.2 Single-dish Calibration of the Amplitude . . . 106
10.2.1 Low opacity approximation and implication for T_{cal} . . . 107
10.2.2 Absolute errors on T_{cal} due to instrumental parameters . . . 110
10.2.3 Relative errors or errors on T_{cal}^L/T_{cal}^U . . . 110
10.2.4 Estimate of the thermal noise . . . 111
10.3 Flux Calibration (visitor's nightmare) . . . 111
10.3.1 Introduction . . . 111
10.3.2 Calibration procedure at Bure . . . 111
10.3.3 Determining the absolute flux scale on a project . . . 113
10.3.4 Possible biases and remedies . . . 115
10.3.5 The program FLUX . . . 115
10.4 Interferometric Calibration of the Amplitude . . . 115
10.4.1 Correction for the antenna gain $\mathcal{B}_i(t)$ . . . 116
10.4.2 Estimate of the atmospheric decorrelation factor f . . . 116
10.4.3 Fitting Splines: the last step . . . 118
10.4.4 A few final checks . . . 119

11 Calibration in Practice **121**
11.1 Introduction . . . 121
11.1.1 Contents of the account . . . 121
11.1.2 Before starting the data reduction . . . 122
11.1.3 Activating the CLIC environment . . . 122
11.2 The "First Look" procedure . . . 122
11.3 The "Standard Calibration (2-receivers)" procedure . . . 123
11.3.1 Inputs . . . 123
11.3.2 Actions or Outputs . . . 124
11.3.3 Results of the calibration . . . 125

12 *UV* Plane Analysis **127**
12.1 *uv* tables . . . 127
12.1.1 *uv* table contents . . . 127
12.1.2 How to create a *uv* Table . . . 128
12.2 *uv* data plots . . . 129
12.3 Data editing . . . 130
12.4 Position shift . . . 131
12.5 Averaging . . . 131
12.5.1 Data compression . . . 131
12.5.2 Circular averaging . . . 131
12.6 Model fitting . . . 132
12.6.1 Position measurement . . . 133
12.7 Continuum source subtraction . . . 133
12.8 Self calibration by a point source . . . 133

13 The Imaging Principles **137**
13.1 Fourier Transform . . . 138
13.1.1 Direct Fourier Transform . . . 138
13.1.2 Fast Fourier Transform . . . 138
13.1.3 Gridding Process . . . 138
13.2 Sampling & Aliasing . . . 139
13.3 Convolution and Aliasing . . . 139
13.4 Error Analysis . . . 141
13.5 Weighing and Tapering . . . 141
13.6 The GILDAS implementation . . . 142
13.7 Deconvolution . . . 143
13.7.1 The CLEAN method . . . 143
13.7.2 Interpretation of CLEAN . . . 144
13.7.3 The CLEAN variants . . . 145
13.7.4 The GILDAS implementation . . . 145

14 Advanced Imaging Methods: WIPE **147**
14.1 Introduction . . . 147
14.2 Object space . . . 148
14.3 Experimental data space . . . 149
14.4 Image reconstruction process . . . 150
14.4.1 Synthesized aperture . . . 150
14.4.2 Synthetic beam . . . 151
14.4.3 Regularization frequency list . . . 151
14.4.4 Data space . . . 152
14.4.5 Object representation space . . . 152
14.4.6 Objective functional . . . 152
14.4.7 Uniqueness and robustness . . . 153
14.5 Implementation of WIPE at IRAM . . . 156

15 Mosaicing **159**
15.1 Introduction . . . 159
15.2 Image formation in a mosaic . . . 160
15.3 Mosaicing in practice . . . 162
15.4 A CLEAN-based algorithm for mosaic deconvolution . . . 164
15.5 Artifacts and instrumental effects . . . 167
15.6 Concluding remarks . . . 170

16 Imaging in Practice **171**
16.1 Visualisation . . . 171
16.2 Photometry . . . 172
16.2.1 From Flux density to Brightness temperature . . . 172
16.2.2 Accuracy of Flux density estimates . . . 172
16.3 Short Spacings . . . 174
16.3.1 UV_SINGLE . . . 174
16.4 Dirty Tricks . . . 175
16.4.1 MOMENTS . . . 175
16.4.2 Continuum Subtraction . . . 176

17 Basic Principles of Radio Astrometry **177**
17.1 Introduction and Basic Formalism . . . 177
17.2 The Phase Equation . . . 178
17.3 Determination of Source Coordinates and Errors . . . 179
17.4 Accurate Position Measurements with the IRAM Interferometer . . . 180
17.4.1 Absolute positions . . . 180
17.4.2 Relative Positions and Self-calibration Techniques . . . 181

Chapter 1

Radio Antennas

Albert Greve

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

1.1 Introduction

We can define a *radio antenna* as an instrument which *collects*, and *detects*, *electromagnetic radiation* from a certain *area* and *direction* of the sky, allowing to build up an image from individual observations. In radio astronomy we are interested in the detection and analysis of radiation emitted from celestial objects, i.e. solar system bodies, stars, interstellar gas, galaxies, and the universe. The electromagnetic radiation observed in radio astronomy covers the wavelength range from several meters, say 10 m (= 30 MHz), to a fraction of a millimeter, say 0.3 mm ($\approx$ 1000 GHz). Since the antenna must be many wavelengths in diameter in order to collect a large amount of energy and to provide a reasonable directivity (angular resolution), it is evident that antennas for meter wavelengths may have dimensions of many 10 meters to several 100 meters, while antennas for millimeter wavelengths have dimensions of several meters to several 10 meters ($\approx$ 10 000 to 50 000 λ's). The technique of mechanical contruction is therefore different for meter and millimeter wavelength antennas: antennas for m–wavelengths can be constructed, for instance, as mesh–wire networks and plate arrays, mm–wavelength antennas are full–aperture solid surface parabolic reflector antennas. Typical examples are the obsolete Mills–Cross antenna, the Effelsberg and GBT 100–m antennas, and the IRAM 30–m (Pico Veleta) and 15–m (Plateau de Bure) antennas. However, despite the variety of mechanical constructions, all antennas can be understood from basic principles of eletromagnetic radiation, optics, and diffraction.

Here we discuss full–aperture parabolic antennas, like the IRAM antennas, which are used for observations at $\sim 3 - 0.8$ mm wavelength (100 – 350 GHz). These antennas are very similar to optical reflector telescopes and use in particular the Cassegrain configuration of a parabolic main reflector and a hyperbolic subreflector (Figure 1.3), with an image formed at the secondary focus near the vertex of the main reflector where the receiver, or a receiver–array, is installed. These antennas are steerable and can observe in any direction of the visible hemisphere, with the facility of tracking, scanning, and mapping of a source.

The collected radiation is concentrated in the secondary focus and is (coherently) detected by a receiver at a certain frequency ν (or wavelength λ) and within a certain bandwidth $\Delta\nu$ (or $\Delta\lambda$). Heterodyne mm–wavelength receivers, which preserve the phase of the incident radiation, have small bandwidths of the

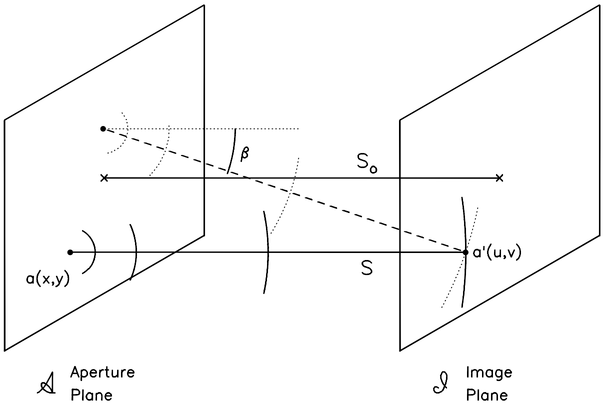


Figure 1.1: Illustration of Huygens Principle. The individual points of the plane wavefront in the aperture plane ($\mathcal{A}$) are the origin of secondary spherical wavelets, which propagate to the right, and superpose to form a plane wavefront in the image plane ($\mathcal{I}$). The optical instrument (telescope) is placed in between $\mathcal{A}$ and $\mathcal{I}$.

order of $\Delta\nu = 0.5 - 2$ GHz so that $\Delta\nu/\nu = \Delta\lambda/\lambda \approx 0.5\,\mathrm{GHz}/100\,\mathrm{GHz} \approx 1/200$. From the point of view of antenna optics, these receivers detect "monochromatic" radiation, and the antenna characteristics can be calculated for a monochromatic wave (as will be done below). Bolometer receivers, on the other hand, detect power in a broad bandwidth of the order of $\Delta\nu \approx 50\,\mathrm{GHz}$ so that $\Delta\nu = 50\,\mathrm{GHz}/250\,\mathrm{GHz} \approx 1/5$. These detectors are no longer monochromatic, and the chromatism of the antenna must be considered in their application.

The construction and operation of a radio antenna is based on exact physical theories, like Maxwell's theory of electromagnetic radiation, the pointing theory of an astronomical instrument, the transformation (mixing, down–conversion, amplification) and detection of electromagnetic radiation, etc. The theory of a radio antenna presented here is, however, only the very tip of an iceberg (of several 100 000 published pages), but may provide sufficient information for the user astronomer to understand the basic principle of a telescope, either a perfect one, which nobody has but which can be described with high precision, or a real one, with small defects and aberrations, which can be described with sufficient detail to apply corrections.

The theory, construction, and use of radio antennas is contained in many textbooks and journals such as IEEE Transactions Antennas and Propagation, Radio Science, Applied Optics. A biased selection is mentioned here: [Born and Wolf 1975] [Reynolds et al, 1989] [Love 1978] [Lo and Lee] [Kraus 1982] [Goldsmith 1988]

1.2 Basic Principles

The properties of electromagnetic radiation propagation and of radio antennas can be deduced from a few basic physical principles, i.e.

1. the notion that Electromagnetic Radiation are Waves of a certain Wavelength (λ), or Frequency (ν), and Amplitude ($\mathbf{A}$) and Phase (φ);

2. from Huygens Principle which says that each element of a wavefront is the origin of a Secondary Spherical Wavelet;

3. the notion that the Optical Instrument (like a single–dish antenna, a telescope, etc.) combined with a receiver manipulates the incident wavefront through their respective phase and amplitude transfer functions.

Summarized in one sentence, and proven in the following, we may say that the radio antenna transforms the radiation incident on the aperture plane ($\mathcal{A}$) to an image in the image plane ($\mathcal{I}$), also called focal plane.

Following Huygens Principle illustrated in Figure 1.1, the point a(x,y) $\equiv$ a($\vec{r}$) of the incident wavefront in the aperture plane $\mathcal{A}$ is the origin of a spherical wavelet of which the field δE(a$'$) at the point a$'$(u,v) $\equiv$ a($\vec{u}$) in the image plane $\mathcal{I}$ is

$$\delta \mathrm{E}(\vec{u}) = \mathrm{A}(\vec{r})\mathrm{exp[iks]/s} \tag{1.1}$$

with $\mathrm{k} = 2\pi/\lambda$. The ensemble of spherical wavelets arriving from all points of $\mathcal{A}$ at the point a$'$($\vec{u}$) of the image plane $\mathcal{I}$ produces the field

$$\mathrm{E}(\vec{u}) = \int_{\mathcal{A}} \mathrm{A}(\vec{r})\Lambda(\beta)[\mathrm{exp(iks)/s}]dxdy \tag{1.2}$$

For the paraxial case, when the rays are not strongly inclined against the direction of wave propagation (i.e. the optical axis), the inclination factor Λ can be neglected since $\Lambda(\beta) \approx \cos(\beta) \approx 1$. Also, $\mathrm{s} \approx \mathrm{s_0}$ for paraxial rays, but exp[iks] $\neq$ exp[iks$_0$] since these are cosine and sine terms of s where a small change in s may produce a large change of the cosine or sine value. Thus, for the paraxial approximation we may write

$$s = [(\mathrm{x-u})^2 + (\mathrm{y-v})^2 + \mathrm{z}^2]^{1/2} \;\approx\; \mathrm{R} + \mathrm{g(x,y,R)} - \mathrm{(xu + yv)/R} \tag{1.3}$$

with

$$\mathrm{R} = (\mathrm{x}^2 + \mathrm{y}^2 + \mathrm{z}^2)^{1/2} \;\; and \;\; \mathrm{g(x,y,R)} = (\mathrm{x}^2 + \mathrm{y}^2)/2\mathrm{R} \tag{1.4}$$

When using these expressions in Eq.1.2, we obtain

$$\mathrm{E(u,v)} = [\mathrm{exp(ikR)/s_0}]\int_{\mathcal{A}} \mathrm{A(x,y)exp[ik(g(x,y,R) - (ux + vy)/R)]dxdy} \tag{1.5}$$

This equation describes the paraxial propagation of a wavefront, for instance the wavefront arriving from a very far away star. In particular, this equation says, that without disturbances or manipulations in between $\mathcal{A}$ and $\mathcal{I}$ the plane wavefront continues to propagate in straight direction as a plane wavefront.

1.3 The perfect Single–Dish antenna

We now place an optical instrument (a mirror, lens, telescope etc.) in the beam between $\mathcal{A}$ and $\mathcal{I}$ with the intention, for instance, to form an image of a star. Optical instruments are invented and developed already since several centuries; however, the physical–optics (diffraction) understanding of the image formation started only a good 200 years ago. Thus, speaking in mathematical terms, the telescope (T) manipulates the phases (not so much the amplitudes) between the points ($\vec{r}$) of the aperture plane ($\mathcal{A}$) and the points ($\vec{u}$) of the image plane ($\mathcal{I}$) by the phase transfer function $\Omega_\mathrm{O}(\vec{r},\vec{u})$, so that the wavefront converges in the focal point. The receiver (R)/detector introduces an additional modulation of the amplitude $\Omega_\mathrm{R}(\vec{r},\vec{u})$, as described below. Using this information, the field distribution in the focal plane ($\mathcal{I}$) of the telescope becomes

$$\mathrm{E}(\vec{u}) = [\mathrm{exp(ikR)/s_0}]\int_{\mathcal{A}} \mathrm{A}(\vec{r})\Omega_\mathrm{O}\Omega_\mathrm{R}\mathrm{exp[ik(g(x,y,R) - (ux + vy)/R)]dxdy} \tag{1.6}$$

The phase modulation of the parabolic reflector used in a radio telescope is, fortunately,

$$\Omega_\mathrm{O} = \mathrm{exp[-ik\,g(x,y,F)]} \tag{1.7}$$

(where F is the focal length of the reflector), which inserted into Eq.1.6 eliminates this term in the exponent so that

$$\mathrm{E}(\vec{u}) = [\mathrm{exp(ikF)/F}]\int_{\mathcal{A}} \mathrm{A}(\vec{r})\Omega_\mathrm{R}\mathrm{exp[-ik(ux + vy)/F]dxdy} \;\equiv\; \mathcal{FT}[\mathrm{A}(\vec{r})\Omega_\mathrm{R}(\vec{r})] \tag{1.8}$$

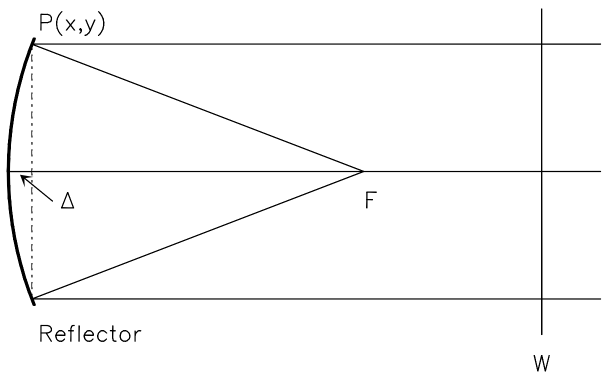


Figure 1.2: Phase modulation of a reflector. The plane wavefront (W) propagates to the left and the ray reflected at P(x,y) toward the focus F is shifted in phase by the amount Δ.

This equation says that the field distribution $\mathrm{E}(\vec{u})$ in the focal plane of the telescope is the Fourier transform ($\mathcal{FT}$) of the receiver–weighted field distribution $\mathrm{A}(\vec{r})\Omega_{\mathrm{R}}(\vec{r})$ in the aperture plane. Since $\mathrm{E}(\vec{u})\mathrm{E}^*(\vec{u}) \neq \delta(\vec{u}-\vec{u}_o)$ for a realistic optical instrument/telescope with limited aperture size, we arrive at the well known empirical fact that the image of a point–like object is not point–like; or; with other words, the image of a star is always blurred by the beam width of the antenna $\Theta_{\mathrm{b}} \propto \lambda/\mathrm{D}$, with D the diameter of the reflector.

To close the argumentation, we need to show that the telescope manipulates the incident wave in the way given by Eq.1.7). To demonstrate this property in an easy way, we consider in Figure 1.2 the paraxial rays of a parabolic reflector of focal length F. From geometrical arguments we have

$$(\mathrm{F}-\Delta)^2 + (\mathrm{x}^2+\mathrm{y}^2) = \mathrm{F}^2 \tag{1.9}$$

which for small Δ becomes

$$\Delta = -(\mathrm{x}^2+\mathrm{y}^2)/\mathrm{F} \equiv -\mathrm{g}(\mathrm{x},\mathrm{y},\mathrm{F}) \equiv \Omega_{\mathrm{O}}(\vec{r}) \tag{1.10}$$

which is the instrumental phase modulation function Ω_{O} used above. The proof is given for a simple parabolic reflector; however, a combined telescope with main reflector and subreflector can be treated in a similar way, leading to the same result.

The fundamental Eq.1.8 can be used to show that an interferometer is *not* a single dish antenna, even though one tries with many individual telescopes and many telescope positions (baselines) to simulate as good as possible the aperture of a large reflector. If we assume for the single dish antennas that $\mathrm{A}(\vec{r}) \equiv 1$ and $\Omega_{\mathrm{R}} \equiv 1$, then the power pattern $\mathrm{P}(\vec{u})$ (beam pattern) in the focal plane of the single antenna is

$$\mathrm{P}(\vec{u}) = \mathrm{E}(\vec{u})\mathrm{E}^*(\vec{u}) = \int_{\mathcal{A}}\int_{\mathcal{A}} \exp[-\mathrm{ik}\vec{u}(\vec{x}_1-\vec{x}_2)](\mathrm{dxdy})_1(\mathrm{dxdy})_2 \ \propto \ [\mathrm{J}_1(\vec{u})/\mathrm{u}]^2 \tag{1.11}$$

where J_1 is the Bessel function of first order (see Born and Wolf, 1980). The function $[\mathrm{J}_1(\mathrm{u})/\mathrm{u}]^2$ is called Airy function, or Airy pattern. The interferometer does not simulate a continuous surface, but consists of individual aperture sections $\mathcal{A}_\infty$, $\mathcal{A}_\in$, ... of the individual telescopes, so that its power pattern $\mathrm{P}_\Sigma(\vec{u})$ (beam pattern) in the focal plane is

$$\mathrm{P}_\Sigma(\vec{u}) = \sum\nolimits_{\mathrm{n}}\sum\nolimits_{\mathrm{m}}\int_{\mathcal{A}\mathrm{n}}\int_{\mathcal{A}\mathrm{m}} \exp[\mathrm{ik}\vec{u}(\vec{x}_1-\vec{x}_2)](\mathrm{dxdy})_1(\mathrm{dxdy})_2 \ \neq \ \mathrm{P}(\vec{u}) \tag{1.12}$$

The important result of this equation is the fact that the image obtained with the interferometer is "incomplete", though certainly not as blurred as seen with a single telescope ($\Theta_{\mathrm{D}} \propto \lambda/\mathrm{D}$), but having the superior resolution of the spatial dimension (approximately the longest baseline B) of the array ($\Theta_{\mathrm{B}} \propto \lambda/\mathrm{B}$). For the Plateau de Bure interferometer $\mathrm{B/D} \approx 300\,\mathrm{m}/15\,\mathrm{m} \approx 20$ so that $\Theta_{\mathrm{B}} \approx 1/20\ \Theta_{\mathrm{D}}$. The incompleteness sometimes requires (in particular for mm–VLBI observations which are very similar) additional information for a full image reconstruction, for instance that the object consists of several point–like sources, or a point–like source and a surrounding halo, etc. (see for instance the number of components in CLEAN).

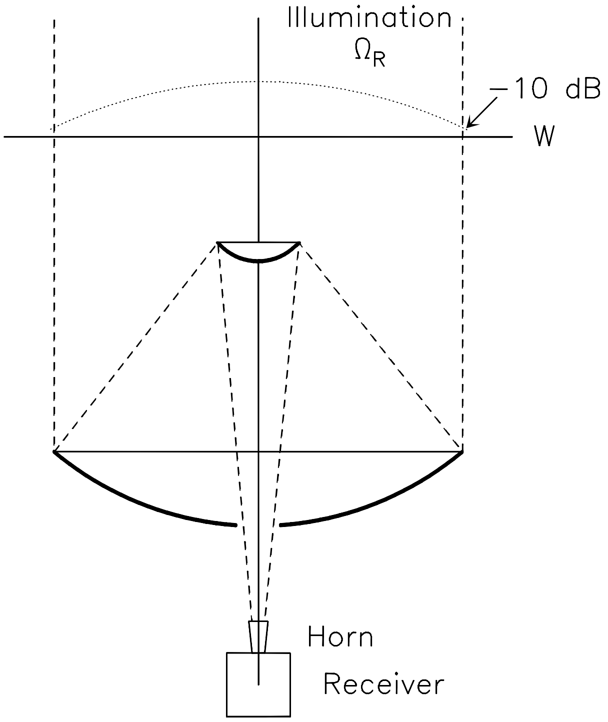


Figure 1.3: The figure shows a Cassegrain telescope and the illumination Ω_R of the incident plane wavefront W.

The single telescope selects a part of the incident plane wavefront and 'bends' this plane into a spherical wave which converges toward the focus. This spherical wavefront enters the receiver where it is mixed, down–converted in frequency, amplified, detected, or correlated. The horn–lens combination of the receiver modifies the amplitude of the spherical wavefront in a way expressed by the function $\Omega_R(\vec{r})$. This function, called taper or illumination function of the horn–lens combination, weighs the wavefront across the aperture, usually in a radial symmetric way. Figure 1.3 shows, schematically, the effect of a parabolic taper as often applied on radio telescopes, and expressed as

$$\Omega_R(\rho) = K + [1-\rho^2]^p \tag{1.13}$$

with ρ the normalized radius of the circular aperture, and K and p being constants. For $A(\vec{r}) \equiv 1$ (i.e. an incident wavefront without structure) the diffraction integral is

$$E_T(\vec{u}) = \int_{\mathcal{A}} \Omega_R(\vec{x}) \exp[-ik\vec{u}\vec{x}] dxdy \quad and \quad E(\vec{u})E^*(\vec{u}) \equiv A_T(\vec{u}) \tag{1.14}$$

E_T is the tapered field distribution in the focal plane, and A_T the tapered beam pattern.

Figure 1.4 shows as example a two–dimensional cut through the calculated beam pattern A_T of the IRAM 15–m telescope at $\lambda = 3\,\mathrm{mm}$, once without taper (i.e. for $\Omega_R(\vec{x}) \equiv 1$), and for a -10 dB edge taper, i.e. when the weighting of the wavefront at the edge of the aperture is 1/10 of that at the center (see Figure 1.3). As seen from the figure, the taper preserves the global structure of the non–tapered beam pattern, i.e. the main beam and side lobes, but changes the width of the beam (BW: Θ_b), the position of the first null (Θ_{fb}), and the level of the side lobes. The effect of the taper depends on the steepness of the main reflector used in the telescope, as shown in Figure 1.5. The influence of several taper forms is given in Table 1.1 [Christiansen and Hogbom 1969].

The complete telescope, i.e. the optics combined with the receiver, has a beam pattern $A_T(\vec{u})$ (in optics called point–spread–function) with which we observe point–like or extended objects in the sky with the intention to know their position, structural detail, and brightness distribution B_S as function of wavelength. The telescope thus provides information of the form

$$I(\vec{u}) \propto \int_{\mathrm{Source}} A_T(\vec{u}-\vec{u}')B_S(\vec{u}')d\vec{u}' \tag{1.15}$$

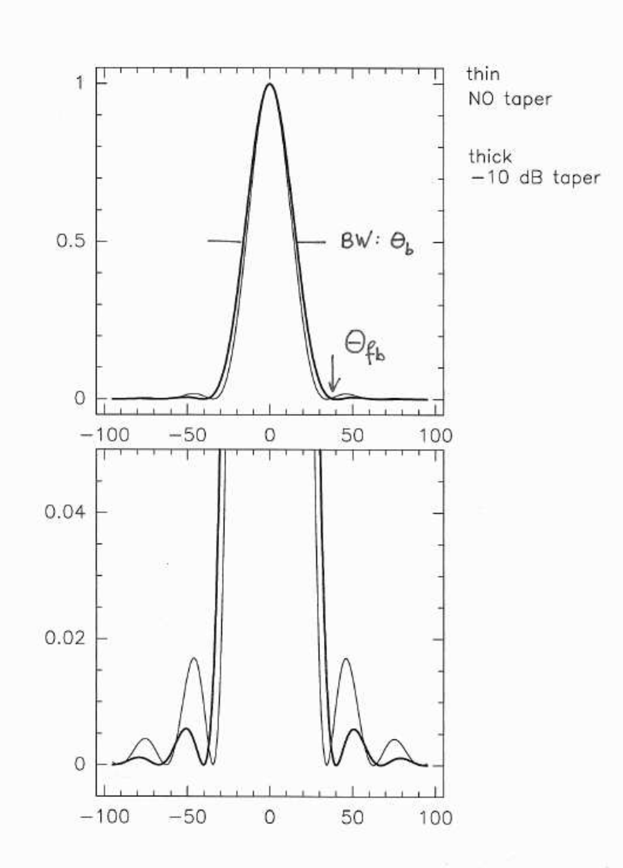


Figure 1.4: The figure shows the calculated beam pattern of the IRAM 15–m telescope at $\lambda = 3$ mm without taper, and for a −10dB edge taper. The horizontal scale is in arcseconds, the vertical scale is normalized power.

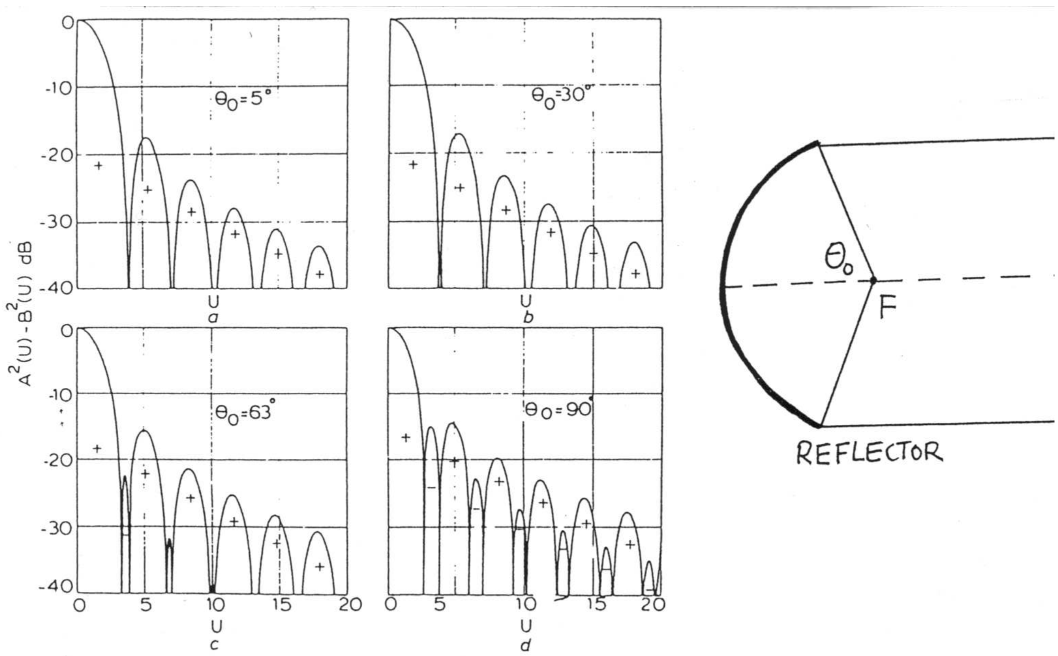


Figure 1.5: Illustration of the tapered beam pattern for telescopes using parabolic main reflectors of different steepness, expressed by the opening angle θ_o. IRAM 30–m and 15–m telescopes: $\theta_o \approx 63°$, optical telescopes: $\theta_o \approx 5°$. Taken from [Minnet and Thomas 1968]

$$\Omega_{\rm R}(x,y) = K + (1-\rho^2)^p$$

$$E(\vec{u}) = \int_A \Omega_{\rm R}(x,y)\exp(-ik\vec{u}\vec{x})dS$$

p	K	Beamwidth Θ_b (radian)	First null Θ_{fb} (radian)	First Sidelobe (dB)	Maximum Aperture Efficiency
0	0	1.02 λ/D	1.22 λ/D	17.6	1.00
1	0	1.27 λ/D	1.62 λ/D	24.7	0.75
2	0	1.47 λ/D	2.03 λ/D	30.7	0.55
1	0.25	1.17 λ/D	1.49 λ/D	23.7	0.87
2	0.25	1.23 λ/D	1.68 λ/D	32.3	0.81
1	0.5	1.13 λ/D	1.33 λ/D	22.0	0.92
2	0.5	1.16 λ/D	1.51 λ/D	26.5	0.88

Table 1.1: Beamwidths, side lobe levels, and maximum aperture efficiency (ϵ_o) for various parameters of the tapering function. From [Christiansen and Hogbom 1969]

If the telescope is perfect, and we know $\rm A_T$, we can use the information $\rm I(\vec{u})$ to derive the calibrated brightness distribution $\rm B_S$ of the source. distribution.

When we point the antenna toward the sky, in essence we point the beam in the direction of observation. If, for instance, we observe a point–like source it is evident that the peak of the main beam should point exactly on the source which requires that the pointing errors ($\Delta\Theta$) of the telescope should be small in comparison to the beam width. The loss in gain is small, and acceptable, if the mispointing $\Delta\Theta < 1/10\,\Theta_{\rm b}$. Since modern radio telescopes use an alt-azimuth mount, this criterion says the mispointing in azimuth ($\Delta\Theta_{\rm az}$) and elevation ($\Delta\Theta_{\rm el}$) direction should not exceed $1/\sqrt{2}$ this value. The pointing and focus (see below) of the IRAM antennas are regularly checked during an observation, and corrected if required. The corresponding protocol of an observing session at Plateau de Bure, using 5 antennas, is shown in Figure 1.6.

1.4 The real Single–Dish Antenna

A telescope, however, is never perfect since mechanical, thermal, and wind–induced deformations of the structure occur, and the optics may be misaligned and/or have production imperfections, for one or the other reason. The resulting effect on the beam pattern is negligible if the corresponding wavefront deformations introduced by these imperfections are small compared to the wavelength of observation, generally smaller than $\sim \lambda/15$; the effect is noticeable and disturbing when the wavefront deformations are large compared to the wavelength ($\sim 1/4\lambda$ and larger). The wavefront deformations due to such imperfections may be of systematic nature, or of random nature, or both.

1.4.1 Sytematic Deformations: Defocus, Coma, Astigmatism

There are three basic systematic surface/wavefront deformations (occasionally associated with pointing errors) with which the observer may be confronted, i.e. defocus, coma, and astigmatism (a transient feature on the IRAM 30–m telescope).

1. The most important systematic wavefront/beam error is due to a **defocus** of the telescope. This error is easily detected, measured, and corrected from the observation of a strong source at a number of focus settings. Figure 1.7 shows, as example, the beam pattern measured on Jupiter with the telescope being gradually defocused. Evidently, the peak power in the main beam decreases, the power in the side lobes increases, until finally the beam pattern has completely collapsed. To be on the safe side for observations, the defocus of the telescope should not exceed $\sim 1/10\lambda$. A defocus does *not* introduce a pointing error.

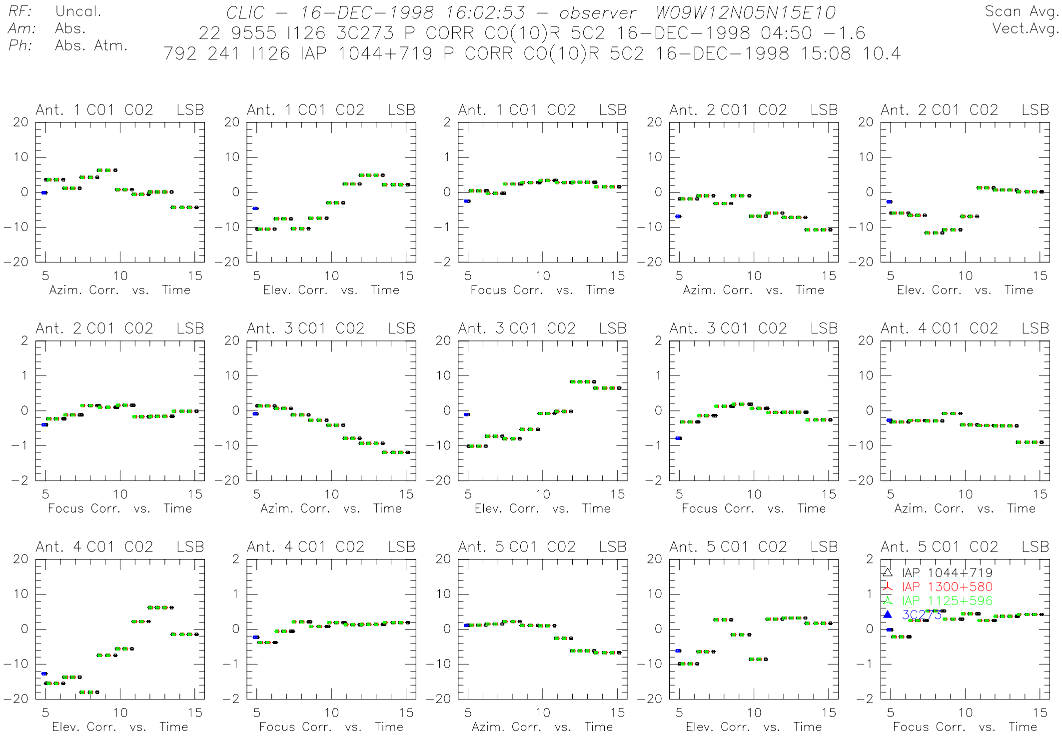


Figure 1.6: Protocol of pointing corrections applied in azimuth and elevation direction, and focus corrections; shown for 5 antennas during an observation which lasted 6 hours.

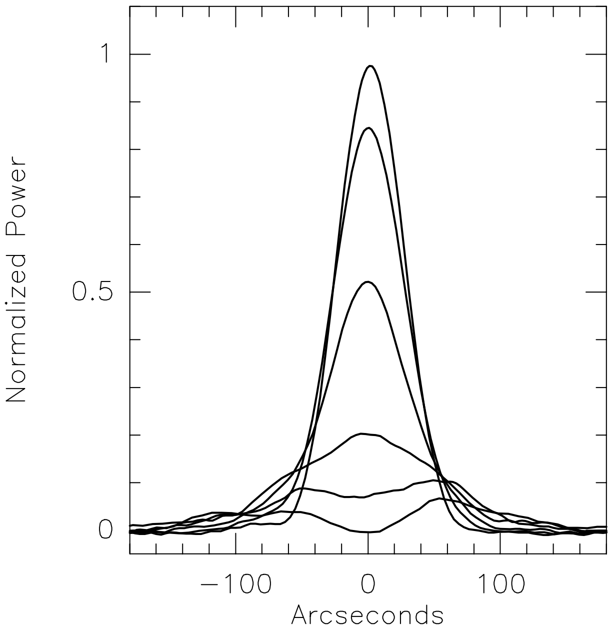


Figure 1.7: Effect on the beam pattern (scans across Jupiter) introduced by defocusing the IRAM 15–m telescope (shifts of the subreflector in steps of $\lambda/4$, $\lambda = 3\,\mathrm{mm}$).

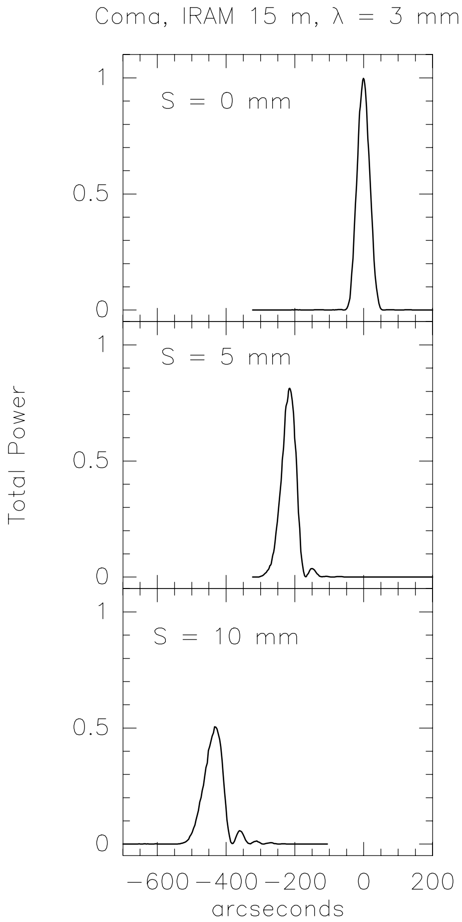


Figure 1.8: Illustration of a comatic beam (scanned in the direction of the coma) especially produced on the IRAM 15–m telescope. The shift of the subreflector is indicated by S. The beam pattern is perfect at S = 0. Note the shift of the beam (pointing error) when the subreflector is shifted.

2. A telescope may have a **comatic** wavefront/beam error due to a misaligned subreflector, shifted perpendicular off the main reflector axis. Figure 8 shows, as example, a cross scan through a comatic beam of the IRAM 15–m telescope, especially produced by displacement of the subreflector. A comatic beam pattern introduces a pointing error. It may be useful for the observer to recognize this error, in particular if unexplained pointing errors occur in an observations. [The IRAM telescopes are regularly checked for misalignments, and correspondingly corrected.]

3. A telescope may have an **astigmatic** wavefront/beam error, usually introduced by complicated mechanical and/or thermal deformations (a transient feature on the IRAM 30–m telescope). While this beam deformation is easily recognized by the observer from the difference in beam widths measured from in–and–out–of–focus cross scans, the improvement of the telescope usually is difficult, and out of reach of the observer. A focused astigmatic beam does *not* introduce a pointing error. Figure 1.9 shows the focused beam pattern measured on a telescope which has a strong astigmatic main reflector (amplitude of the astigmatism $\sim 0.5\,\mathrm{mm}$).

The beam deformation of systematic wavefront deformations occurs close to the main beam, and the exact analysis should be based on diffraction calculations. A convenient description of systematic deformations uses Zernike polynomials of order (n,m) [Born and Wolf 1975]. Without going into details, the Zernike–type surface deformation $\delta_{\rm n,m} = \alpha_{\rm n,m}\, R_{\rm n}(\rho)\, \cos(m\theta)$ [with (ρ, θ) normalized coordinates of the aperture, and R special polynomial functions] with amplitude $\alpha_{\rm n,m}$ has a quasi rms–value $\sigma = \alpha_{\rm n,m}/\sqrt{\rm n+1}$ and introduces a loss in main beam intensity of

$$\epsilon_{\rm sys}/\epsilon_o \approx \exp[-(4\pi\alpha/\lambda)^2/({\rm n}+1)] \qquad (1.16)$$

For primary coma n = 1, for primary astigmatism n = 2. Although the beam deformation may be very noticeable and severe, the associated loss in main beam intensity may still be low because of the reduction

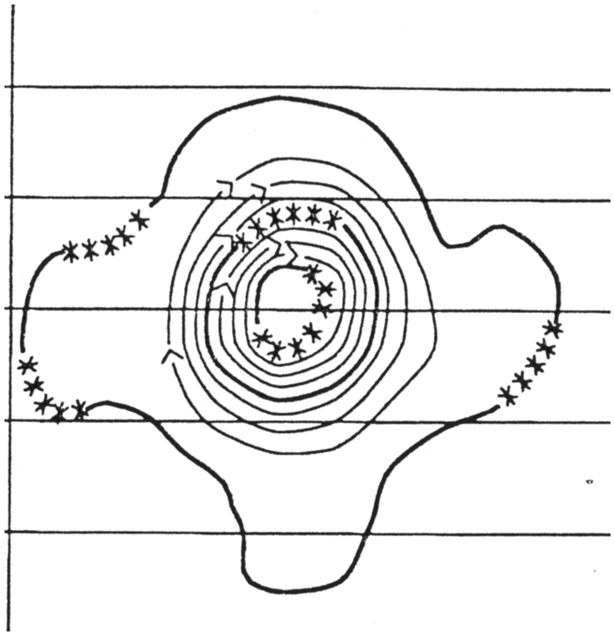


Figure 1.9: Illustration of an astigmatic beam pattern; well focused.

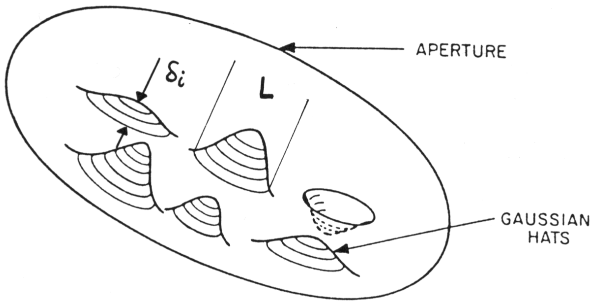


Figure 1.10: Explanation of random errors (δ) and their correlation length (L), for Gaussian hat–like deformations (example). [Ruze 1996]

by the factor (n+1).

1.4.2 Random Errors

Besides systematic surface/wavefront deformations explained above (mainly due to misalignment of the optics), there are often permanent random deformations on the optic surfaces like ripples, scratches, dents, twists, misaligned panels, etc., with spatial dimensions ranging from several wavelengths to significant areas of the aperture. These deformations introduce identical deformations of the wavefront, which cannot be expressed in mathematical form (as the Zernike polynomials used above). Nevertheless, the effect on the beam pattern of this type of deformations can be analyzed in a statistical way and from a simple expression, the RUZE equation. This equation is often used to estimate the quality of a telescope, in particular as function of wavelength. The values obtained from this equation are directly related to the aperture efficiency, and beam efficiency, of the telescope, and hence are important for radiometric measurements (see Sect.1.5).

As illustrated in Figure 1.10, there are two parameters which allow a physical–optics description of the influence of random errors, i.e. the rms–value (root mean square value) σ of the deformations, and their correlation length L.

Random errors occur primarily on the main reflector; the other optical components of the telescope (subreflector, Nasmyth mirror, lenses, polarizers) are relatively small and can be manufactured with good precision. In order to explain the rms–value σ, we assume that the reflector aperture is divided into *many* elements (i = 1,2,...N), and that for each element [i] the deformation δ(i) of the reflector is known with respect to a smooth mean surface. The rms–value of these random surface deformations is

$$\sigma = \sqrt{\sum_{\mathrm{i=1,N}} \delta(\mathrm{i})^2/\mathrm{N}} \tag{1.17}$$

The surface deformations $\delta(\mathrm{i})$ introduce corresponding wavefront deformations $\varphi(\mathrm{i})$, approximately two times larger than the mechanical deformations δ in case we are dealing with reflective optics. The rms–value σ_φ of the corresponding phase deformations of the wavefront is

$$\sigma_\varphi = 2\,\mathrm{k}\,\mathrm{R}\,\sigma \tag{1.18}$$

again with $\mathrm{k} = 2\pi/\lambda$, and $\mathrm{R} \approx 0.8$ a factor which takes into account the steepness of the parabolic main reflector [Greve & Hooghoudt 1981].

A description of the wavefront deformation by the rms–value σ_φ is incomplete since the value does not contain information on the structure of the deformations, for instance whether they consist of many dents at one part of the aperture, or many scratches at another part. A useful physical–optics description requires also a knowledge of the correlation length L of the deformations. L is a number ($\mathrm{L} \le \mathrm{D}$) which quantifies the extent over which the randomness of the deformations does not change. For example, the deformations of a main reflector constructed from many individual panels, which may be misaligned, often has a random error correlation length typical of the panel size, but also a correlation length of 1/3 to 1/5 of the panel size due to inaccuracies in the fabrication of the individual panels. A typical example is the 30–m telescope [Greve et al 1998].

When knowing, by one or the other method, the rms–value σ_φ and the correlation length L, it is possible to express the resulting beam shape in an analytic form which describes well the real situation. The beam pattern $\mathcal{F}(\Theta)$ of a wavefront with random deformations $(\sigma_\varphi, \mathrm{L})$ [the telescope may actually have several random error distributions] consists of the degraded diffraction beam $\mathcal{F}_\mathrm{c}(\Theta)$ and the error beam $\mathcal{F}_\mathrm{e}(\Theta)$ such that

$$\mathcal{F}(\Theta) = \mathcal{F}_\mathrm{c}(\Theta) + \mathcal{F}_\mathrm{e}(\Theta) \tag{1.19}$$

with

$$\mathcal{F}_\mathrm{c}(\Theta) = \exp[\,-\,(\sigma_\varphi)^2]\mathrm{A_T}(\Theta) \tag{1.20}$$

where $\mathrm{A_T}(\Theta)$ is the tapered beam pattern (Eq.(5)), and

$$\mathcal{F}_\mathrm{e}(\Theta) = \mathrm{a}\exp[-(\pi\Theta\mathrm{L}/\lambda)^2] \tag{1.21}$$

where

$$\mathrm{a} = (\mathrm{L/D})^2[1 - \exp(-\sigma_\varphi^2)]/\epsilon_o \tag{1.22}$$

In these equations is D the diameter of the telescope aperture, λ the wavelength of observation, Θ the angular distance from the beam axis, and ϵ_o the aperture efficiency of the perfect telescope. In the formalism used here the beam is circular symmetric. The error beam $\mathcal{F}_\mathrm{e}(\Theta)$ has a Gaussian profile of width FWHP) $\Theta_\mathrm{e} = 0.53\lambda/\mathrm{L}$ [radians], i.e. the smaller the correlation length (the finer the irregular structure), the broader is the beam width Θ_e. The random errors of panel surface deformation and panel alignment errors may have large error beams (up to arcminutes in extent) which can pick up radiation from a large area outside the actual source. A knowledge of the structure and of the level of the error beam(s) is therefore important when mapping a source and making absolute power measurements. Figure 1.11 shows the diffraction beam and the combined error patterns measured on the 30–m telescope at various wavelengths. The smaller the wavelength of observation, the smaller is the power received in the main beam and the larger the power received in the error beam. Due to its particular mechanical construction, this telescope has three error beams [Greve et al 1998].

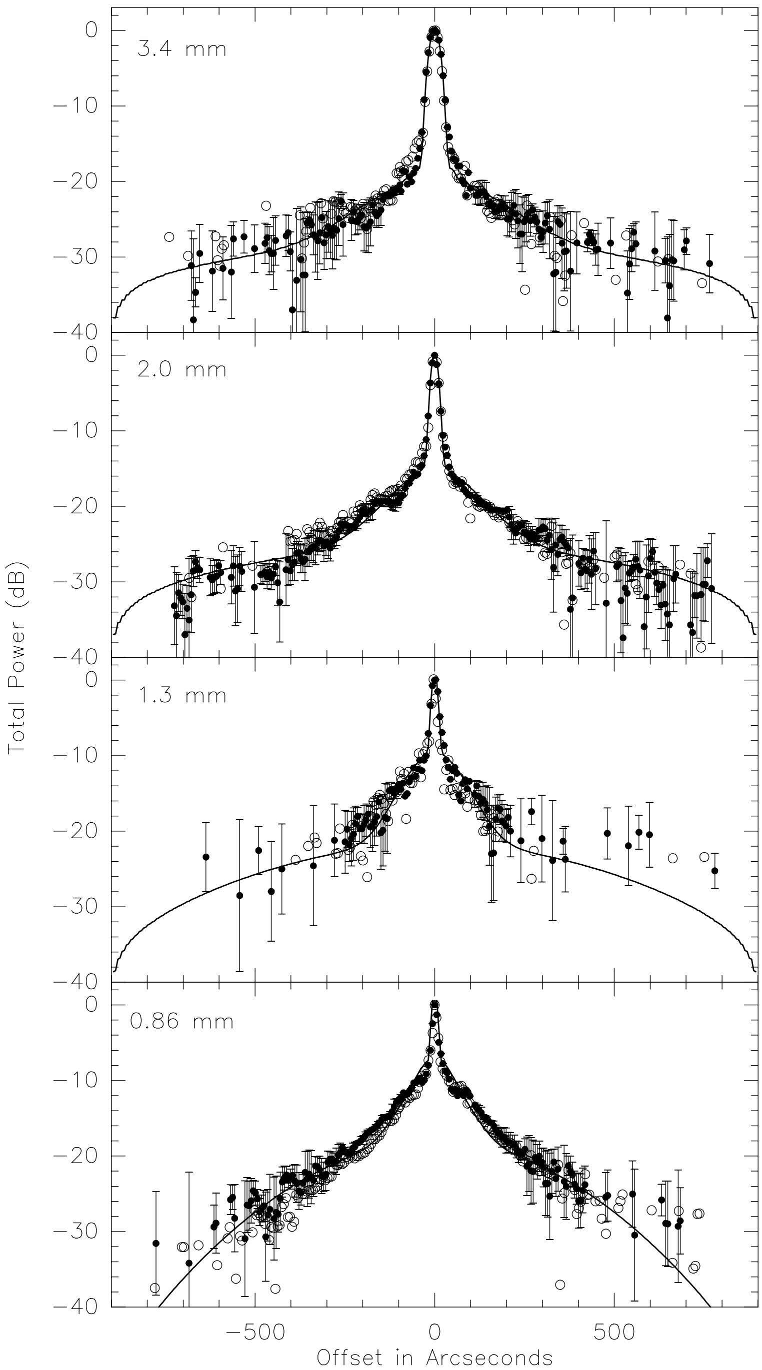


Figure 1.11: Beam pattern measured on the IRAM 30–m telescope. The beam consists of the diffraction beam (≈ main beam) and a combined, extended error beam (solid line).

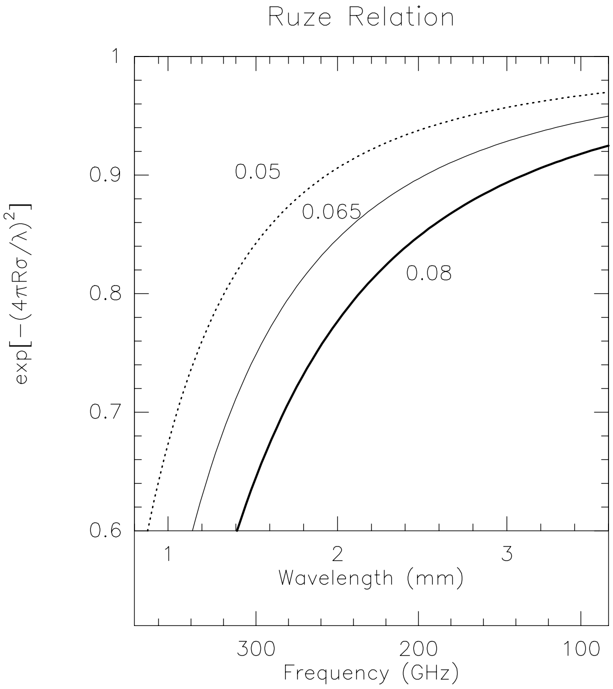


Figure 1.12: Illustration of the RUZE relation exp[–$(4\pi\mathrm{R}\sigma/\lambda)^2$] as function of wavelength (frequency) and values Rσ as indicated (mm).

1.5 Radiometric Relations

The imperfections of a telescope, either due to systematic or random errors, produce beam deformations, a loss in gain, and focus and pointing errors. These effects must be taken into account when mapping and measuring a source. Information on the beam pattern obtained from a map (for instance holography map) of a strong point–like source; information on the sensitivity [Jy/K] and calibration of the telescope is obtained from absolute power measurements of, for instance, the planets, of which the brightness temperatures are quite well known. This information is usually collected by the observatory staff, and provided to the observer (30–m Telescope Manual; observation protocols of Plateau de Bure measurements).

We summarize the influence of random deformations, at least as far as the main beam is concerned, since for this case the RUZE equation provides sufficient precision for an understanding of the telescope behaviour; also for the astronomer observer without going into complicated radio optics detail. This relation appears in the expression of the diffraction beam $\mathcal{F}_\mathrm{c}$ (see Eq.1.20) and shows clearly the fact that the degradation of the telescope, in particular for power measurements, increases exponentially with wavelength.

Aperture Efficiency:

$$\epsilon_\mathrm{ap} = \epsilon_o \,\mathbf{exp}[-\sigma_\varphi] = \epsilon_o \mathbf{exp}[-(4\pi\mathrm{R}\sigma/\lambda)^2]$$

Antenna Gain:

$$\mathrm{S}/\mathrm{T}_\mathrm{A}^* = 2(\mathrm{k}/\mathrm{A})\mathrm{F}_\mathrm{eff}/\epsilon_\mathrm{ap} \approx\; 2(\mathrm{k}/\mathrm{A})\,\mathbf{exp}[+(4\sigma\mathrm{R}/\lambda)^2]/\epsilon_\mathrm{ap} \quad [\mathrm{Jy/K}]$$

Beam Efficiency:

$$\mathrm{B}_\mathrm{eff} = 0.8899\,[\Theta_\mathrm{b}/(\lambda/\mathrm{D})]^2/\epsilon_\mathrm{ap}$$

$$\Theta_\mathrm{b} = \alpha\lambda/\mathrm{D}, \quad 1 \leq \alpha\ 1.2 \quad [\mathrm{radians}]$$

$$\mathrm{B}_\mathrm{eff} \approx 1.2\,\epsilon_o \mathbf{exp}[-(4\pi\mathrm{R}\sigma/\lambda)^2]$$

The quantities in these expressions are ϵ_o: aperture efficiency of the perfect telescope (usually of the order of $\sim 75 - 90\,\%$; see Table 1.1); ϵ_ap: effective aperture efficiency at the wavelength λ, including all wavefront / telescope deformations; σ: rms–value of the telescope optics deformations; R $\approx$ 0.8: reduction factor for a steep main reflector (N = F/D $\approx$ 0.3); S: flux density of a point source [Jy]; T_A^*: measured antenna temperature [K]; A: geometric surface area of the telescope [m^2]; F_eff: forward efficiency, measured at the telescope for instance from a sky dip; Θ_b: main beam width (FWHP).

Chapter 2

The interferometer principles

Stéphane Guilloteau

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

This lecture presents the principle of the heterodyne interferometer. An heterodyne interferometer is composed of `antennas` (described in A.Greve lecture 1), `receivers` (described in B.Lazareff lecture 3), a `correlator` (described in H.Wiesemeyer lecture 4) and an awful lot of cables and connections. This lecture only describes the basic principles; a more complete description, including subtleties due to multiple frequency conversions and digital delay lines, is given in R.Lucas lecture 5.

2.1 Basic principle

The `antenna` produces a Voltage proportional to the linear superposition of the incident electric field pattern. For a simple monochromatic case:

$$U(t) = E\cos(2\pi\nu t + \Phi) \tag{2.1}$$

In the `receiver`, a `mixer` superimposes the field generated by a `local oscillator` to the antenna output.

$$U_{LO}(t) = Q\cos(2\pi\nu_{LO} t + \Phi_{LO}) \tag{2.2}$$

The mixer is a non-linear element (diode) whose output is

$$I(t) = a_0 + a_1(V(t) + V_{LO}(t)) + a_2(V(t) + V_{LO}(t))^2 + a_3(V(t) + V_{LO}(t))^3 + ... \tag{2.3}$$

The second order (quadratic) term of Eq.3 can be expressed as

$$\begin{aligned} I(t) = \quad & ... \\ + \quad & a_2 E^2 \cos^2(2\pi\nu t + \Phi) \\ + \quad & 2a_2 EQ \cos(2\pi\nu t + \Phi)\cos(2\pi\nu_{LO} t + \Phi_{LO}) \\ + \quad & a_2 Q^2 \cos^2(2\pi\nu_{LO} t + \Phi_{LO}) \\ + \quad & ... \end{aligned} \tag{2.4}$$

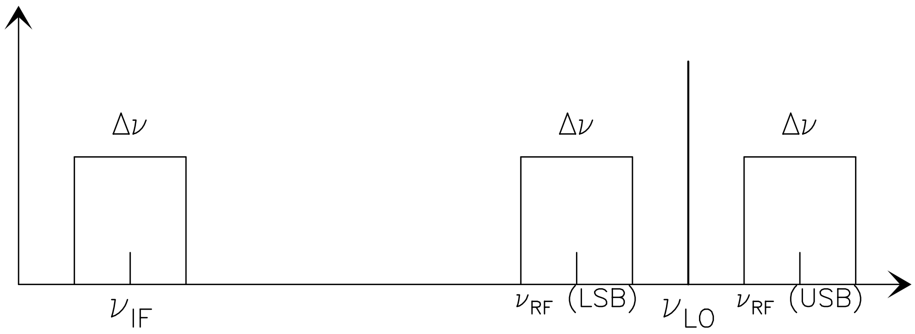


Figure 2.1: Relation between the IF, RF and local oscillator frequencies in an heterodyne system

Developping the product of the two cosine functions, we obtain

$$\begin{aligned} I(t) = \quad & ... \\ + \quad & a_2 EQ\cos(2\pi(\nu+\nu_{LO})t + \Phi + \Phi_{LO}) \\ + \quad & a_2 EQ\cos(2\pi(\nu-\nu_{LO})t + \Phi - \Phi_{LO}) \\ + \quad & ... \end{aligned} \tag{2.5}$$

There are obviously other terms in $2\nu_{LO}$, 2ν, $3\nu_{LO} \pm \nu$, etc ... in the above equation, as well as terms at very different frequencies like ν, 3ν, etc ...
By inserting a filter at the output of the `mixer`, we can select only the term such that

$$\nu_{IF} - \Delta\nu/2 \leq |\nu - \nu_{LO}| \leq \nu_{IF} + \Delta\nu/2 \tag{2.6}$$

where ν_{IF}, the so-called *Intermediate Frequency*, is a frequency which is significantly different from than the original signal frequency ν (which is often called the *Radio Frequency* ν_{RF}).
Hence, after mixing and filtering, the output of the `receiver` is

$$I(t) \propto EQ\cos(2\pi(\nu - \nu_{LO})t + \Phi - \Phi_{LO}) \tag{2.7}$$

or

$$I(t) \propto EQ\cos(2\pi(\nu_{LO} - \nu)t - \Phi + \Phi_{LO}) \tag{2.8}$$

i.e.

- changed in frequency: $\nu \to \nu - \nu_{LO}$ or $\nu \to \nu_{LO} - \nu$
- proportional to the original electric field of the incident wave: $\propto E$
- with a phase relation with this electric field: $\Phi \to \Phi - \Phi_{LO}$ or $\Phi \to \Phi_{LO} - \Phi$
- proportional to the `local oscillator` voltage: $\propto Q$

The frequency change, usually towards a lower frequency, allows to select ν_{IF} such that amplifiers and transport elements are easily available for further processing. The mixer described above accepts simultaneously frequencies which are (see Fig.2.1)

- higher than the `local oscillator` frequency.
  This is called Upper Side Band (USB) reception
- lower than the `local oscillator` frequency.
  This is called Lower Side Band (LSB) reception

and cannot *a priori* distinguish between them. This is called Double Side Band (DSB) reception. Some receivers are actually insensitive to one of the frequency range, either because a filter has been placed at the receiver input, or because their response is very strongly frequency dependent. Such receivers are called Single Side Band (SSB) receivers.

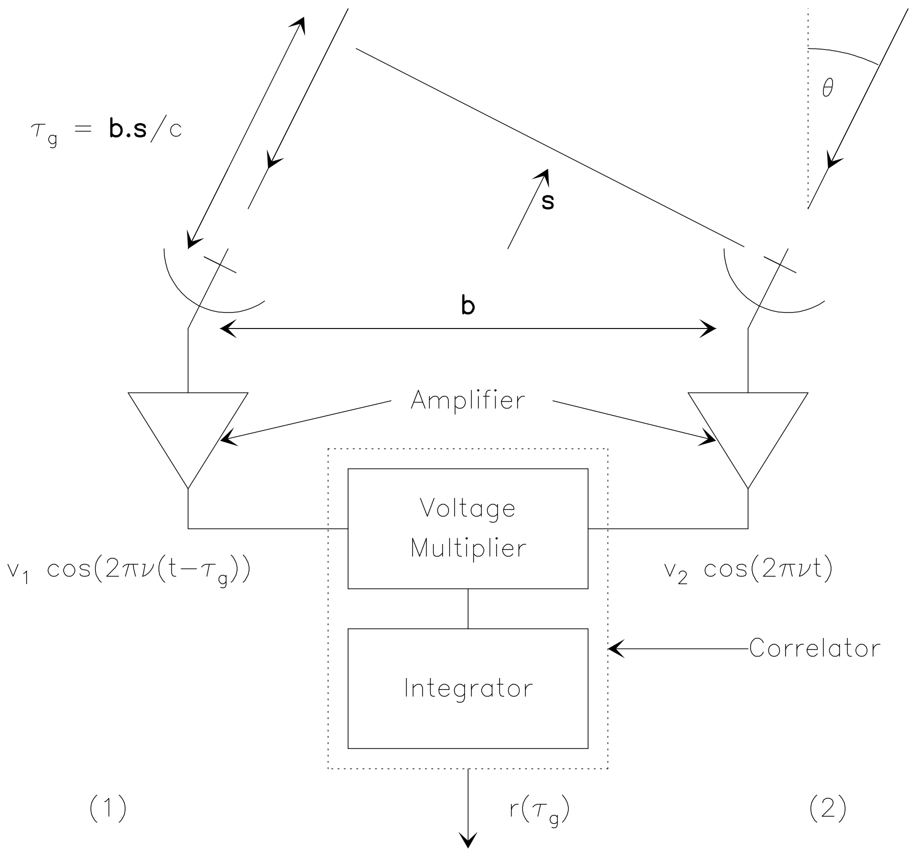


Figure 2.2: Schematic Diagram of a two-element interferometer

An important property of the receiving system expressed by Eq.2.8 is that the sign of the phase is changed for LSB conversion. This property can be easily retrieved recognizing that the **Frequency ν is the time derivative of the Phase Φ**. Assume the phase varies linearly with time:

$$\begin{aligned} \Phi(t) &= 2\pi n t \\ n &= \frac{1}{2\pi}\frac{d\Phi}{dt} \end{aligned} \tag{2.9}$$

In this case, the signal

$$\begin{aligned} I(t) &\propto \cos(2\pi\nu t + \Phi(t)) \\ &\propto \cos(2\pi(\nu + n)t) \end{aligned} \tag{2.10}$$

is just another monochromatic signal with slightly shifted frequency.

2.2 The Heterodyne Interferometer

Figure 2.2 is a schematic illustration of a 2-antenna heterodyne interferometer.

Let us forget the frequency conversion for some time, i.e. assume $\nu_{IF} = \nu_{RF}$...

The input (amplified) signals from 2 elements of the interferometer are processed by a `correlator`, which is just a voltage multiplier followed by a time integrator. With one incident plane wave, the output $r(t)$ is

$$r(t) =< v_1 \cos(2\pi\nu(t - \tau_g(t)))v_2 \cos(2\pi\nu t) >= v_1 v_2 \cos(2\pi\nu\tau_g(t)) \tag{2.11}$$

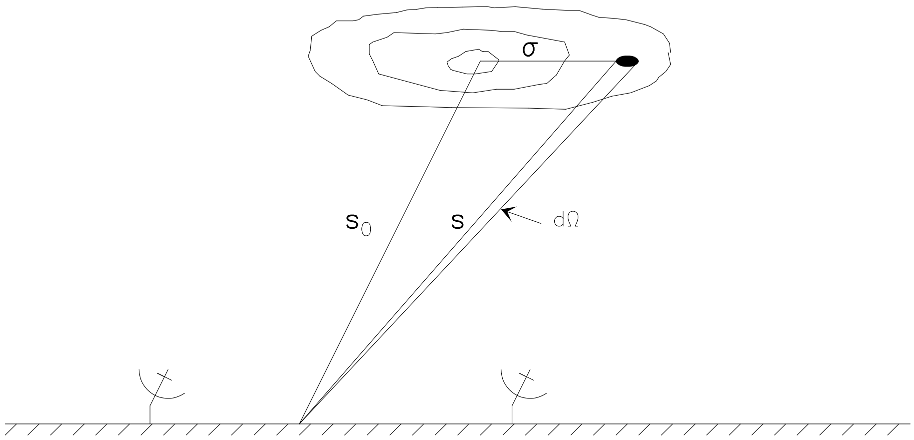


Figure 2.3: Position vectors used for the expression of the interferometer response to an extended source, schematically represented by the iso-contours of the sky brightness distribution.

where τ_g is obviously the geometrical delay,

$$\tau_g(t) = (\mathbf{b}.\mathbf{s})/c \tag{2.12}$$

As τ_g varies slowly because of Earth rotation, $r(t)$ oscillates as a cosine function, and is thus called the *fringe pattern*. As we had shown before that v_1 and v_2 were proportional to the electric field of the incident wave, the correlator output (fringe pattern) is thus proportional to the power of the wave.

2.2.1 Source Size Effects

The signal power received from a sky area $d\Omega$ in direction $\mathbf{s}$ is (see Fig.2.3 for notations) $A(\mathbf{s})I(\mathbf{s})d\Omega d\nu$ over bandwidth $d\nu$, where $A(\mathbf{s})$ is the antenna power pattern (assumed identical for both elements, more precisely $A(\mathbf{s}) = A_i(\mathbf{s})A_j(\mathbf{s})$ with A_i the voltage pattern of antenna i), and $I(\mathbf{s})$ is the sky brightness distribution

$$dr = A(\mathbf{s})I(\mathbf{s})d\Omega d\nu \cos(2\pi\nu\tau_g) \tag{2.13}$$

$$r = d\nu \int_{Sky} A(\mathbf{s})I(\mathbf{s})\cos(2\pi\nu\mathbf{b}.\mathbf{s}/c)d\Omega \tag{2.14}$$

Two implicit assumptions have been made in deriving Eq.2.14. We assumed incident plane waves, which implies that the source must be in the far field of the interferometer. We used a linear superposition of the incident waves, which implies that the source must be spatially incoherent. These assumptions are quite valid for most astronomical sources, but may be violated under special circumstances. For example VLBI observations of solar system objects would violate the first assumption, while observations of celestial masers could (in theory) violate the second one.

When the interferometer is tracking a source in direction $\mathbf{s}_o$, with $\mathbf{s} = \mathbf{s}_o + \sigma$

$$\begin{aligned} r &= d\nu\cos(2\pi\nu\mathbf{b}.\mathbf{s}_o/c)\int_{Sky} A(\sigma)I(\sigma)\cos(2\pi\nu\mathbf{b}.\sigma/c)d\Omega \\ &- d\nu\sin(2\pi\nu\mathbf{b}.\mathbf{s}_o/c)\int_{Sky} A(\sigma)I(\sigma)\sin(2\pi\nu\mathbf{b}.\sigma/c)d\Omega \end{aligned} \tag{2.15}$$

We define the *Complex Visibility*

$$V = |V|e^{i\Phi_V} = \int_{Sky} A(\sigma)I(\sigma)e^{(-2i\pi\nu\mathbf{b}.\sigma/c)}d\Omega \tag{2.16}$$

which resembles a Fourier Transform...

We thus have

$$\begin{aligned} r &= d\nu\left(\cos(2\pi\nu\mathbf{b}.\mathbf{s}_o/c)|V|\cos(\Phi_V) - \sin(2\pi\nu\mathbf{b}.\mathbf{s}_o/c)|V|\sin(\Phi_V)\right) \\ &= d\nu|V|\cos(2\pi\nu\tau_g - \Phi_V) \end{aligned} \tag{2.17}$$

i.e. the `correlator` output is proportional to the amplitude of the visibility, and contains a phase relation with the visibility.

2.2.2 Finite Bandwidth

Integrating over $d\nu$,

$$R = \frac{1}{\Delta\nu}\int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2} |V|\cos(2\pi\nu\tau_g - \Phi_V)d\nu \tag{2.18}$$

Using $\nu = \nu_0 + n$

$$R = \frac{1}{\Delta\nu}\int_{-\Delta\nu/2}^{\Delta\nu/2} |V|\cos(2\pi\nu_0\tau_g - \Phi_V + 2\pi n\tau_g)dn \tag{2.19}$$

$$\begin{aligned} &= \frac{1}{\Delta\nu}[\int_{-\Delta\nu/2}^{\Delta\nu/2} |V|\cos(2\pi\nu_0\tau_g - \Phi_V)\cos(2\pi n\tau_g)dn \\ &\quad - \int_{-\Delta\nu/2}^{\Delta\nu/2} |V|\sin(2\pi\nu_0\tau_g - \Phi_V)\sin(2\pi n\tau_g)dn] \end{aligned} \tag{2.20}$$

$$\begin{aligned} &= \frac{1}{\Delta\nu}|V|\cos(2\pi\nu_0\tau_g - \Phi_V)\,[\sin(2\pi n\tau_g)]_{-\Delta\nu/2}^{\Delta\nu/2}\frac{1}{2\pi\tau_g} \\ &\quad + \frac{1}{\Delta\nu}|V|\sin(2\pi\nu_0\tau_g - \Phi_V)\,[\cos(2\pi n\tau_g)]_{-\Delta\nu/2}^{\Delta\nu/2}\frac{1}{2\pi\tau_g} \end{aligned} \tag{2.21}$$

$$= |V|\cos(2\pi\nu_0\tau_g - \Phi_V)\frac{\sin(\pi\Delta\nu\tau_g)}{\pi\Delta\nu\tau_g} \tag{2.22}$$

The fringe visibility is attenuated by a $\sin(x)/x$ envelope, called the bandwidth pattern, which falls off rapidly. A 1% loss in visibility is obtained for $|\Delta\nu\tau_g| \simeq 0.078$, or with $\Delta\nu = 500$MHz and a baseline length $b = 100$m, when the zenith angle θ (defined in Fig.2.3) is 2 arcmin only. Thus, the ability to track a source for a significant hour angle coverage requires proper compensation of the geometrical delay when a finite bandwidth is desired.

2.3 Delay Tracking and Frequency Conversion

To compensate for the geometrical delay variations, delay lines with mirrors (as in optics...) would be completely impractical given the required size of the mirrors. The compensating delay is thus performed electronically after one (or several) frequency conversion(s), as illustrated in Fig.2.4. This can be implemented either by switching cables with different lengths, or in a more sophisticated way, by using shift memories after digital sampling of the signal in the correlator. Apart for a few details (see R.Lucas lecture), the principle remains identical.

In the case presented in Fig.2.4, for USB conversion, the phase changes of the input signals before reaching the correlator are

$$\Phi_1 = 2\pi\nu\tau_g = 2\pi(\nu_{LO} + \nu_{IF})\tau_g \tag{2.23}$$

$$\Phi_2 = 2\pi\nu\tau_i + \Phi_{LO} \tag{2.24}$$

Introducing $\Delta\tau = \tau_g - \tau_i$ as the delay tracking error, the correlator output is

$$\begin{aligned} r &= A_o|V|\cos(\Phi_1 - \Phi_2 - \Phi_V) \\ USB\quad r &= A_o|V|\cos(2\pi(\nu_{LO}\tau_g + \nu_{IF}\Delta\tau) - \Phi_V - \Phi_{LO}) \\ LSB\quad r &= A_o|V|\cos(2\pi(\nu_{LO}\tau_g - \nu_{IF}\Delta\tau) - \Phi_V - \Phi_{LO}) \end{aligned} \tag{2.25}$$

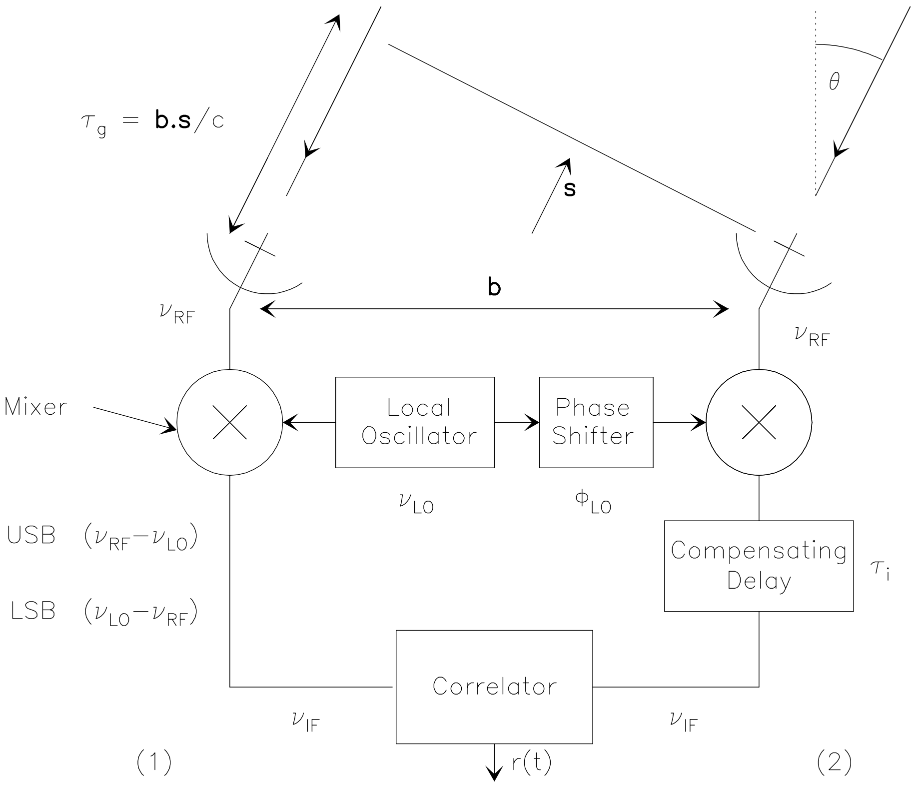


Figure 2.4: 2-element heterodyne interferometer with delay tracking after frequency conversion

When the two sidebands are superposed,

$$DSB \quad r \quad = \quad 2A_o|V|\cos(2\pi(\nu_{LO}\tau_g - \Phi_V - \Phi_{LO}))\cos(2\pi\nu_{IF}\Delta\tau) \tag{2.26}$$

i.e. the amplitude is modulated by the delay tracking error. The tolerance can be exceedingly small. For example, at Plateau de Bure, the IF frequency ν_{IF} is 3 GHz, and a 1 % loss is obtained as soon as the delay tracking error would be 7.5 picoseconds, i.e. a geometrical shift of 2.2 mm only. Due to Earth rotation, the geometrical delay changes by such an amount in 0.1 s for a 300 m baseline. Hence, delay tracking would have to be done quite fast to avoid sensitivity losses. To avoid this problem, it is common to use sideband separation. The delay tracking error should then be kept small compared to the bandwidth of each spectral channel, $\tau_g << 1/\Delta\nu$, and the delay can then be adjusted much less frequently.

2.4 Fringe Stopping and Complex Correlator

With the Earth rotation, the cosine term of Eq.2.22 modulates the correlator output quasi-sinusoidaly with a *natural fringe rate* of

$$\nu_{LO} d\tau_g/dt \quad \simeq \quad \Omega_{earth}\mathrm{b}\nu_{LO}/c \tag{2.27}$$

which is of order of 10 Hz for b= 300 m baselines and $\nu_{LO} = 100$ GHz, or $2''$ angular resolution (since the fringe rate only depends on the effective angular resolution).

The fringe rate is somewhat too high for simple digital sampling of the visibility. An exception is VLBI (because there is no other choice), although the resolutions are < 1 mas. The usual technique is to modulate the phase of the `local oscillator` Φ_{LO} such that $\Phi_{LO}(t) = 2\pi\nu_{LO}\tau_g(t)$ at any given time. Then

$$r_r = A_o|V|\cos(\pm 2\pi\nu_{IF}\Delta\tau - \Phi_V) \tag{2.28}$$

(with the + sign for USB conversion, and the – sign for LSB conversion), is a slowly varying output, which would be constant for a point source at the reference position (or delay tracking center). This process is called *Fringe Stopping*. After fringe stopping, we can no longer measure the amplitude $|V|$ and the phase Φ_V separately. A second correlator, with one signal phase shifted by $\pi/2$ becomes necessary. Its output is

$$r_i = A_o|V|\sin(\pm 2\pi\nu_{IF}\Delta\tau - \Phi_V) \tag{2.29}$$

With both correlators, we measure directly the real r_r and imaginary r_i parts of the complex visibility r. The device is thus called a **"complex" correlator**.

Note: A delay tracking error $\Delta\tau$ appears as a phase slope as a function of frequency, with

$$\Phi(\nu_{IF}) = \pm 2\pi\nu_{IF}\Delta\tau \tag{2.30}$$

2.5 Fourier Transform and Related Approximations

The *Complex Visibility* is

$$V = |V|e^{i\Phi_V} = \int_{Sky} A(\sigma)I(\sigma)e^{(-2i\pi\nu\mathbf{b}.\sigma/c)}d\Omega \tag{2.31}$$

Let (u, v, w) be the coordinate of the baseline vector, in units of the observing wavelength ν, in a frame of the phase tracking vector $\mathbf{s}_0$, with w along $\mathbf{s}_0$. (x, y, z) are the coordinates of the source vector $\mathbf{s}$ in this frame. Then

$$\begin{aligned} \nu\mathbf{b}.\mathbf{s}/c &= ux + vy + wz \\ \nu\mathbf{b}.\mathbf{s}_0/c &= w \\ z &= \sqrt{1-x^2-y^2} \\ \text{and} \quad d\Omega &= \frac{dxdy}{z} = \frac{dxdy}{\sqrt{1-x^2-y^2}} \end{aligned} \tag{2.32}$$

Thus,

$$V(u, v, w) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} A(x, y)I(x, y)e^{-2i\pi(ux+vy+w(\sqrt{1-x^2-y^2}-1))}\frac{dxdy}{\sqrt{1-x^2-y^2}} \tag{2.33}$$

with $I(x, y) = 0$ when $x^2 + y^2 \geq 1$.

If (x, y) are sufficiently small, we can make the approximation

$$(\sqrt{1-x^2-y^2} - 1)w \simeq \frac{1}{2}(x^2 + y^2)w \simeq 0 \tag{2.34}$$

and Eq.2.33 becomes

$$V(x, y) = \iint A'(x, y)I(x, y)e^{-2i\pi(ux+vy)}dxdy \tag{2.35}$$

$$\text{with} \quad A'(x, y) = \frac{A(x, y)}{\sqrt{1-x^2-y^2}} \tag{2.36}$$

i.e. basically a 2-D Fourier Transform of AI, but with a phase error term $\pi(x^2 + y^2)w$. Hence, on limited field of views, the relationship between the sky brightness (multiplied by the antenna power pattern) and the visibility reduces to a simple 2-D Fourier transform.

There are other approximations related to field of view limitations. Let us quantify these approximations.

- **2-D Fourier Transform**
  We can further neglect the phase error term in Eq.2.36, if the condition
  $$|\pi(x^2+y^2)w| << 1 \tag{2.37}$$
  is fulfilled. Now, note that
  $$w < w_{\max} \simeq \frac{b_{\max}}{\lambda} \simeq \frac{1}{\theta_{\mathrm{s}}} \tag{2.38}$$
  where *Beams* is the synthesized beam width. Thus, if θ_{f} is the field of view to be synthesized, the maximum phase error is
  $$\Delta\phi = \frac{\pi\theta_{\mathrm{f}}^2}{4\theta_{\mathrm{s}}} \tag{2.39}$$
  Using $\Delta\phi < 0.1$ radian (6°) as an upper limit (note that this is the maximum phase error, i.e. the mean phase error is much smaller) result in the condition (with all angles in radian...):
  $$\theta_{\mathrm{f}} < \frac{1}{3}\sqrt{\theta_{\mathrm{s}}} \tag{2.40}$$

- **Bandwidth Smearing**
  Assume u, v are computed for the center frequency ν_0. At frequency ν_0, we have
  $$V(u,v) \rightleftharpoons AI(x,y) \tag{2.41}$$
  The similarity theorem on Fourier pairs give
  $$V(\frac{\nu_0}{\nu}u, \frac{\nu_0}{\nu}v) = (\frac{\nu}{\nu_0})^2 I(\frac{\nu}{\nu_0}x, \frac{\nu}{\nu_0}y) \tag{2.42}$$
  Averaging over the bandwidth $\Delta\nu$, there is a *radial smearing* equal to
  $$\sim \frac{\Delta\nu}{\nu_0}\sqrt{x^2+y^2} \tag{2.43}$$
  and hence the constraint
  $$\sqrt{x^2+y^2} \leq 0.1\frac{\theta_{\mathrm{s}}\nu_0}{\Delta\nu} \tag{2.44}$$
  if we want that smearing to be less than 10% of the synthesized beam.

- **Time Averaging**
  Assume for simplicity that the interferometer observes the Celestial Pole. The baselines cover a sector of angular width $\Omega_e\Delta t$, where Ω_e is the Earth rotation speed, and Δt the integration time. The smearing is *circumferential* and of magnitude $\Omega_e\Delta t\sqrt{x^2+y^2}$, hence the constraint
  $$\sqrt{x^2+y^2} \leq 0.1\frac{\theta_{\mathrm{s}}}{\Omega_e\Delta t} \tag{2.45}$$
  For other declinations, the smearing is no longer rotational, but of similar magnitude.

To better fix the importance of such approximations, the relevant values for the Plateau de Bure interferometer are given in Table 2.1. Note that these fields of view correspond to a maximum phase error of 6° only, or to a (one dimensional) distortion of a tenth of the synthesized beam, and thus are not strict limits. In particular, atmospheric errors often results in larger errors (which are independent of the field of view, however).

2.6 Array Geometry & Baseline Measurements

Config.	Resolution	Frequency (GHz)	2-D Field	0.5 GHz Bandwidth	1 Min Time Averaging	Primary Beam
Compact	$5''$	80 GHz	$5'$	$80''$	$2'$	$60''$
Standard	$2''$	80 GHz	$3.5'$	$30''$	$45''$	$60''$
Standard	$2''$	220 GHz	$3.5'$	$1.5'$	$45''$	$24''$
High	$0.5''$	230 GHz	$1.7'$	$22''$	$12''$	$24''$

Table 2.1: Field of view limitations as function of angular resolution and observing frequency for the Plateau de Bure interferometer.

uv **coverage** Using a Cartesian coordinate system (X, Y, Z) with Z towards the pole, X towards the meridian, and Y towards East, the conversion matrix to u, v, w is

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} \sin(h) & \cos(h) & 0 \\ -\sin(\delta)\cos(h) & \sin(\delta)\cos(h) & \cos(\delta) \\ \cos(\delta)\cos(h) & -\cos(\delta)\sin(h) & \sin(\delta) \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \quad (2.46)$$

where h, δ are the hour angle and declination of the phase tracking center.
Eliminating h from Eq.2.46 gives the equation of an ellipse:

$$u^2 + \left(\frac{v - (Z/\lambda)\cos(\delta)}{\sin(\delta)}\right)^2 = \frac{X^2 + Y^2}{\lambda^2} \quad (2.47)$$

The *uv* coverage is an ensemble of such ellipses. The choice of antenna configurations is made to cover the *uv* plane as much as possible.

Baseline measurement Assume there is a small baseline error, $(\Delta X, \Delta Y, \Delta Z)$. The phase error is

$$\Delta\phi = \frac{2\pi}{\lambda}\Delta\mathbf{b}.\mathbf{s}_0 \quad (2.48)$$
$$= \cos(\delta)\cos(h)\Delta X - \cos(\delta)\sin(h)\Delta Y + \sin(\delta)\Delta Z \quad (2.49)$$

Hence, if we observe N sources, we have for each source

$$\phi_k = \phi_0 + \cos(\delta_k)\cos(h_k)\Delta X - \cos(\delta_k)\sin(h_k)\Delta Y + \sin(\delta_k)\Delta Z \quad (2.50)$$

i.e. a linear system in $(\Delta X, \Delta Y, \Delta Z)$, with N equations and 4 unknown (including the arbitrary phase ϕ_0). This can be used to determine the baselines from phases measured on a set of sources with known positions h_k, δ_k.

From the shape of Eq.2.49, one can see that the determination of $\Delta X, \Delta Y$ requires large variations in h, preferably at declination $\delta \sim 0$, while that of ΔZ requires large variations in δ. However, ϕ_k in Eq.2.50 is multi-valued (the 2π ambiguity...). Hence, the system to solve is in fact

$$mod(\phi_0 + \cos(\delta_k)\cos(h_k)\Delta X - \cos(\delta_k)\sin(h_k)\Delta Y + \sin(\delta_k)\Delta Z - \phi_k, 2\pi) = 0 \quad (2.51)$$

which is a linear system of equations only if $\Delta X, \Delta Y, \Delta Z$ are small enough so that the modulo function is the identity. Baseline determination usually proceeds through a "brute force" technique, by making a grid search around the most likely values for X, Y, Z.

Chapter 3

Receivers : an overview for non-specialists

Bernard Lazareff

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

3.1 Introduction

The purpose of a receiver is to collect efficiently the astronomical signal that has been concentrated by the antenna near its focal point, and to amplify it with a minimum of extra noise to a level suitable for further processing by the spectrometers or continuum detectors. Figure 3.1 shows the main subsystems of a receiver, that we will discuss below.

3.2 Coupling optics

Up to and including the antenna, the astronomical signal propagates in free space. On the other hand, the first signal processing unit —the mixer— requires the electromagnetic energy to be confined by metallic walls, in a waveguide. The transition between these two modes of propagation occurs at the *horn*.

Assume for a moment that the horn would be placed at the Cassegrain focus of the antenna. Good matching would be difficult to achieve because the field amplitude from a point source (Airy pattern) exhibits radial oscillations alternating between positive and negative values, and has a scalesize proportional to wavelength.

These problems are avoided by coupling —via suitable relay optics— the horn to an *image of the aperture*. This fulfills the condition of *frequency-independent illumination*. In other words, imagine that we propagate the horn mode back to the antenna aperture as if we were dealing with a transmitter, then the illumination pattern is independent of frequency. Using suitably designed corrugations on the inner wall of the horn (see Fig. 3.2) the TE10 mode of the rectangular waveguide couples to a mode at the aperture of the horn whose amplitude has circular symmetry, and whose polarization is pure linear.

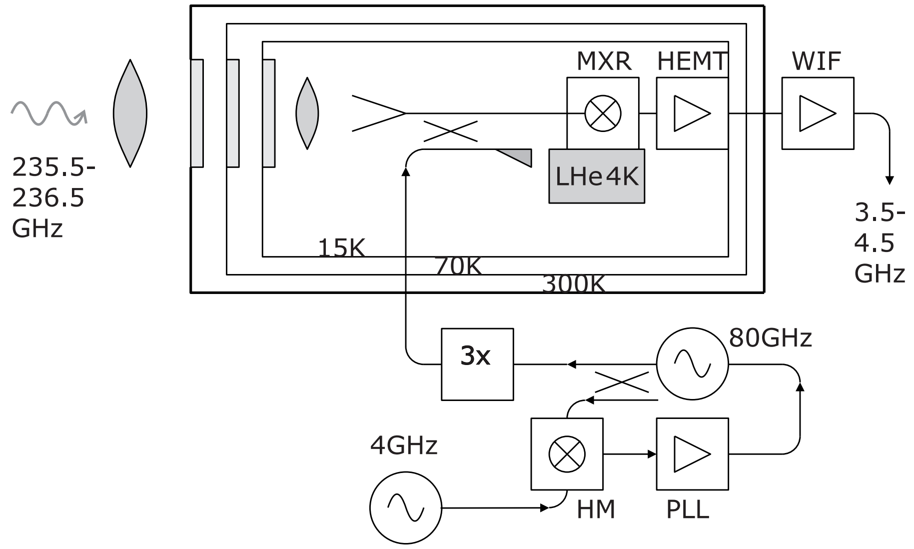


Figure 3.1: Synoptic diagram of a typical receiver. This diagram is grossly simplified; for instance, the optics involves mirrors, elliptical and planar, and also grids; the LO/PLL system is actually more complicated than shown.

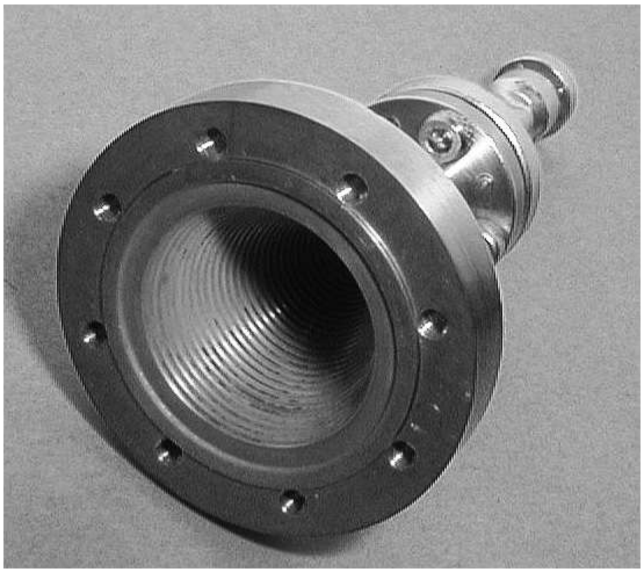


Figure 3.2: A corrugated horn for the 150 GHz band. The phase-correcting lens normally present at the aperture has been removed to reveal the corrugations. The diameter of the aperture is about that of a typical coin (1 DM - 100 Ptas - 1 F). All the waves collected by the 30-m antenna converge to the horn with a precise phase relationship and are then squeezed into a waveguide 2 by 1 mm across.

3.3 Why we need heterodyne receivers

In the present context, *heterodyne* refers to receivers where the frequency of the input signal is shifted to lower frequencies. This is done by adding to the (small) input signal a (relatively) strong monochromatic signal, called the *local oscillator* and passing the sum through a non-linear device, whose output contains (among other) the difference frequency. Although a non-linear device is involved, the transformation from input to output is linear for the small signal (see S.Guilloteau lecture, Eq. 2.8). This process is called mixing or downconversion. The output signal is called the *intermediate frequency.* Actually the complete signal processing at a radiotelescope can involve up to four heterodyne conversions.

The first reason why heterodyne downconversion is needed is that only few signal processing devices exist at millimeter frequencies, and definitely not the fully parallel spectrometers (as opposed to multiplex devices such as FTS) that are routinely used for spectroscopic observations.

Then arises the question of where in the signal processing chain to operate the down conversion. Basically we have no choice, because hardly any amplifiers are available in the millimeter range, except in the 3mm band, where they do not match the low noise properties of SIS mixers (to be discussed below). So we *must* perform a downconversion before we can amplify the signal.

3.4 Local oscillator system

All the local oscillators in the IRAM telescope use Gunn oscillators. A Gunn diode is a semiconductor device that exhibits negative dynamic resistance over a suitable range of frequencies. Output powers of the order of 10–50mW can be obtained between 60 and 120 GHz. To achieve oscillation at a precise frequency, two means are combined. First, the Gunn diode is coupled to a coaxial cavity that defines the oscillation frequency, and whose high quality factor provides a good spectral purity. Its resonant frequency can be adjusted mechanically; this allows the desired frequency to be approached within ≈10 MHz. Secondly, a fraction of the millimetric radiation from the Gunn oscillator is used to produce a beat with a reference microwave oscillator at a frequency of a few GHz; actually, the Gunn oscillator signal beats with a harmonic (n=17–65, depending on the systems) of the reference frequency. The beat signal is used to "servo" by electronic tuning the Gunn oscillator to a multiple of the reference. Actually, not only the frequency, but also the phase of the local oscillator is locked to the reference oscillator, which is essential for interferometry, whether connected-array or VLBI. This description of the *phase-lock* system is over-simplified.

Local oscillator frequencies above 120 GHz can generally not be generated directly by Gunn oscillators. In that case, the Gunn power is fed to a *frequency multiplier*, which is a non-linear device like the mixer, but based on non-linear capacitance, and optimized to produce a certain harmonic (×2, ×3 or ×4 in the case of IRAM systems) of the input frequency. The efficiency of the multiplication process is typically a few percent.

3.5 Local oscillator injection

As mentioned above, the local oscillator power must be added to the astronomical signal before it enters the mixer. When the mixers were based on Schottky diodes (10 years ago and more), they required an LO power of almost a mW. As a consequence, the LO power was coupled via a *diplexer*, which is like a frequency-selective coupler, allowing the mixer to be coupled with close to unity efficiency to *both* the input signal and the LO. With the advent of SIS mixers, and due to their modest LO power requirements (see Eq. 3.5), a new method can be used, based on frequency-independent couplers. A fraction $\mathbf{f}$ (typically 1%) of the LO power is coupled into the signal path; the rest is wasted! The fraction $\mathbf{f}$ must be kept small because the same amount $\mathbf{f}$ of room-temperature blackbody noise is also coupled into the signal path. The coupler requires no adjustment and is located close to the receiver, inside the *cryostat* (see Sec.3.8).

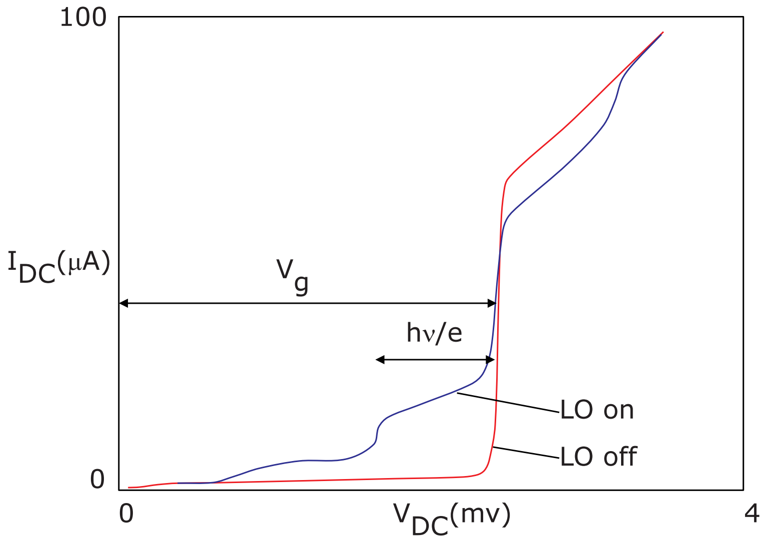


Figure 3.3: Current-voltage characteristics of an SIS junction operating in a mixer. The two curves were measured without and with LO power applied (frequency 230GHz); they have been slightly idealized (for pedagogical reasons, of course).

3.6 Photon-assisted tunneling

All mixers in IRAM receivers are based on SIS junctions. An SIS junction consists of two layers of superconducting metal (Niobium) separated by a few nanometers of insulator (Aluminum oxide). The insulator is so thin that charged particles can tunnel through the barrier. The area of a junction is typically one to a few μm^2. SIS junctions operate at the boiling temperature of He: 4.2K (at sea level), or even below.

Two kinds of charged particles can exist in a superconductor: a) ordinary electrons; b) so-called Cooper pairs, consisting of two electrons interacting and weakly bound together by the exchange of phonons (lattice vibrations); breaking a Cooper pair costs an energy 2Δ. Correspondingly, two kinds of currents can flow across the junction: the Josephson current, consisting of Cooper pairs, and the so-called quasi-particle current, consisting of "ordinary" electrons (presumably "electron" did not sound fancy enough). To keep this digression into SIS physics short, let's just state that the Josephson current can be ignored. At the operating temperature of the mixer, and in an unbiased junction, the population of quasi-particles is virtually negligible. But, if the bias voltage is raised to the *gap voltage*

$$V_g = 2\Delta/e \tag{3.1}$$

the flow of quasi-particles across the junction becomes possible because the energy gained across the drop of electrical potential compensates for the energy spent in breaking a Cooper pair. See on figure 3.3 the "LO off" I-V characteristic.

In the presence of electromagnetic radiation, the situation is modified as follows. If a RF photon is absorbed, its energy $h\nu$ can contribute to the energetic budget, which can now be written as:

$$eV_{\rm bias} + h\nu = 2\Delta \tag{3.2}$$

or, equivalently:

$$V_{\rm bias} + h\nu/e = V_g \tag{3.3}$$

In other words, the onset of conduction occurs at $V_g - h\nu/e$. The region of the I-V curve below the gap voltage where photon-assisted tunneling occurs is called the *photon step*. See the "LO on" curve on Fig. 3.3. Fig. 3.3 is based on actual measurements of a 2-junction series array: the voltage scale has been scaled $\times\frac{1}{2}$ to illustrate a single junction. For a more detailed analysis of SIS junctions and their interaction with radiation (see e.g. [Gundlach 1989]).

So far I have shown you qualitatively that an SIS junction can function as a total power detector. The responsivity (current generated per power absorbed) can even be estimated to be of the order of one electron per photon, or: $D \approx e/h\nu$. How does that relate to frequency downconversion? Assume that a power detector is fed the sum of a local oscillator (normalized to unit amplitude for convenience) $v_{LO} = \cos\omega_{LO}t$ and a much smaller signal at a nearby frequency: $v_S = \epsilon\cos\omega_S t$. Assumes this functions as a squaring device and discard high-frequency terms in the output:

$$v_{\rm out} = (v_{LO} + v_S)^2 \longrightarrow \frac{1}{2}\epsilon\, cos(\omega_{LO} - \omega_S)t \tag{3.4}$$

So, a power detector can also function as a frequency downconverter (subject to possible limitations in the response time of the output).

The LO power requirement for an SIS mixer can be estimated as follows. A voltage scale is defined by the width of the photon step: $h\nu/e$. Likewise, a resistance scale can be defined from R_N, the resistance of the junction above V_g; junctions used in mixers have $R_N \approx 50\,\Omega$. So, the order of magnitude of the LO power required is:

$$P_{\rm LO} \approx (h\nu/e)^2/R_N \tag{3.5}$$

about 20 nW for a 230 GHz mixer. This makes it possible to use the wasteful coupler injection scheme discussed above.

Because the insulating barrier of the junction is so thin, it possesses a capacitance of about $65fF\,\mu m^{-2}$. At the RF and LO frequencies, the (imaginary) admittance of that capacitance is about 3–4× the (approximately real) admittance of the SIS junction itself. Therefore, appropriate tuning structures must be implemented to achieve a good impedance match (i.e. energy coupling) of the junction to the signals.

The minimum theoretical SSB noise for an SIS mixer is $h\nu/k$, 11 K at 230 GHz; the best IRAM mixers come within a factor of a few ($\approx 3\times$) of that fundamental limit. These numbers are for laboratory measurements with minimal optics losses; practical receivers have a slightly higher noise.

3.7 Mixer

A sketch of a mixer is shown on Fig. 3.4, again grossly over-simplified. The junction is mounted across the waveguide, in the direction of the electric field. One side of the junction is connected to the outside of the mixer block, both to bring out the IF beat signal, and to provide the DC bias. That connection is made through a low-pass filter to avoid losing precious RF energy.

One end of the waveguide is the input of the mixer; the other end must be terminated somehow. At the zero-order approximation, one would like the junction to "see" an open circuit when "looking into" the rear end of the waveguide. More generally, the junction should see a pure imaginary impedance, so that no energy is wasted. A simple calculation shows that a transmission line having a length l, and terminated into a short-circuit, has an apparent impedance:

$$Z_{BS} = j\, Z_0 \tan(2\pi l/\lambda) \tag{3.6}$$

where Z_0 and λ are respectively the propagation impedance and wavelength in the waveguide, and l is the distance to the short-circuit. In particular, for $l = (\frac{n}{2} + \frac{1}{4})\lambda$, the apparent impedance is an open circuit. More generally, by adjusting l, an arbitrary imaginary impedance can be placed in parallel with the junction. Together with the tuning structures mentioned in the previous section, such an *adjustable backshort* contributes to achieve the best possible match of the junction impedance.

For various reasons (one of which is reducing the noise contribution from the atmosphere) it is desirable that the mixer should operate in *single-sideband* mode. We explain how this is achieved with a crude zero-order model. Assume that the best impedance match of the junction is obtained when the apparent impedance of the backshort seen from the junction is an open circuit. Assume we observe in the lower sideband at a frequency $\nu_L = \nu_{LO} - \nu_{IF}$, and want to reject the upper sideband $\nu_U = \nu_{LO} + \nu_{IF}$. That

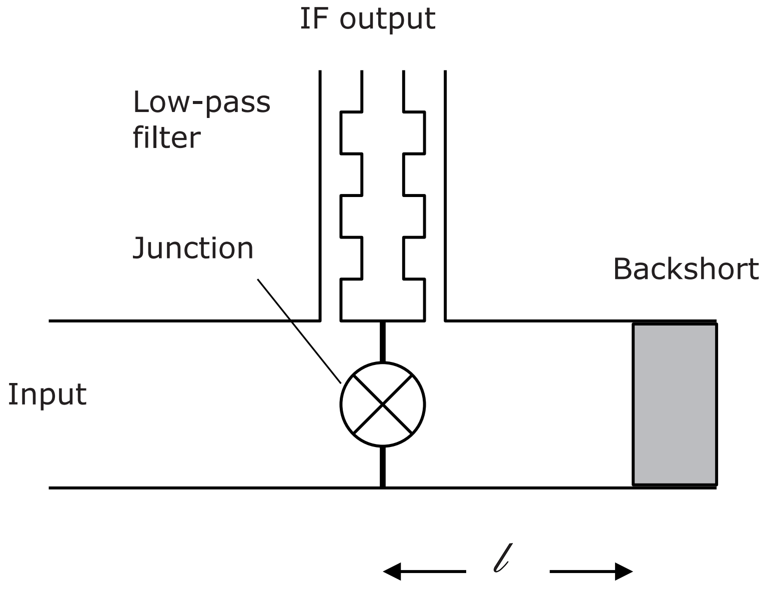


Figure 3.4: Rough sketch of the main elements of a mixer

condition can be achieved if, at the frequency ν_U, the junction is short-circuited. So, we must meet the two conditions:

$$l = (\frac{n}{2} + \frac{1}{4})\lambda_L \quad \text{lower sideband}$$
$$l = (\frac{n}{2} + \frac{1}{2})\lambda_U \quad \text{upper sideband}$$

for some integer n; we gloss over the distinction between free-space and waveguide wavelengths. The two conditions (one unknown) can be *approximately* met for some l close to

$$l_{\text{reject}} = \frac{1}{8}\frac{c}{\nu_{IF}}(\approx 9 \text{ mm for a 4 GHz IF}) \tag{3.7}$$

Because for current mixers in the 100 GHz band, the IF frequency is relatively low (1.5 GHz), single-sideband operation requires additional tricks ...

Returning to practicalities, tuning a receiver requires several steps (which used to make astronomers a bit nervous at the 30-m telescope when all was done manually). First the local oscillator must be tuned and locked at the desired frequency. Then the backshort is set at the appropriate position, and the junction DC bias voltage is set. Finally the LO power is adjusted to reach a prescribed junction DC current (of the order of $20\mu A$). These adjustments are made by a combination of table lookup and optimization algorithms under computer control. Altogether this involves between 11 and 13 adjustments, mechanical or electrical, yet this process takes only a few minutes with the current systems.

3.8 Cryostat

As mentioned earlier, the SIS junctions in the mixers operate at the boiling temperature of He. Therefore, at the heart of the cryostat lies a reservoir of ≈4 liters of liquid He. However, if that would be exposed to ambient conditions, several undesirable things would happen. First, conducted heat would quickly evaporate the helium. Second, a big icicle of water, nitrogen, oxygen, etc ... would condense around the reservoir. Conduction and condensation are avoided by operating the receiver in a vacuum enclosure (labeled 300 K in Fig. 3.1). But infrared radiation must also be blocked. Your body is receiving about

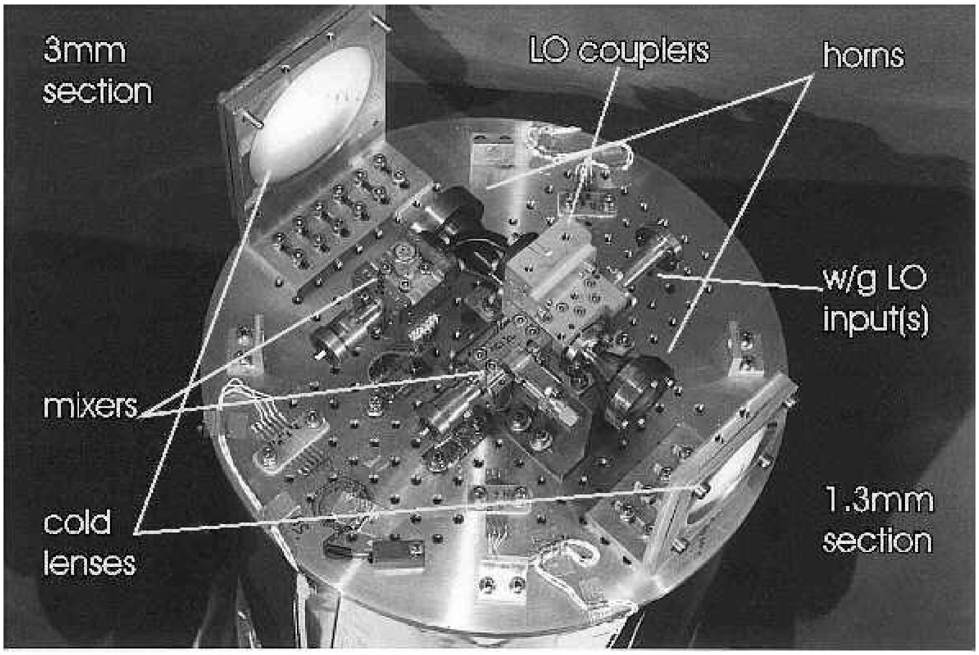


Figure 3.5: The cold RF assembly for a dual-channel receiver used at the 30-m telescope.

700 W from the surroundings! (and radiating back about the same amount). A typical 4-liter reservoir of liquid He, exposed to the same flux, would evaporate in 2 minutes! Yet the hold time of a cryostat is one to several weeks, four orders of magnitude more. This is achieved via the two radiation screens labeled 70 K 15 K in Fig. 3.1, as well as by reducing to a minimum all conduction losses. The radiation screens are kept cold by a closed-cycle cryogenic machine involving the compression and expansion of helium gas. The 15 K stage is also used to cool the first stages of IF amplification. Future receivers will feature fully closed-cycle cryogenics, including the 4 K stage.

3.9 Actual receivers

Fig. 3.5 gives you a chance to peek at the cold RF assembly of one of the dual-channel receivers in operation at the 30-m telescope since May 1998.

Fig. 3.6 shows the performance of one of the Plateau de Bure 230 GHz receivers. The present LO/IF system dictates a 1.5 GHz IF, therefore, these receivers are operated in DSB mode; in the interferometer, the sidebands can be separated due to their different fringe rates. These receivers are among the best currently produced (either at IRAM or elsewhere in the world). Fig. 3.7 shows the performance obtained over a wider frequency band. Noise levels of about 4 times the DSB quantum limit (i.e. $2k\nu/k$) are obtained.

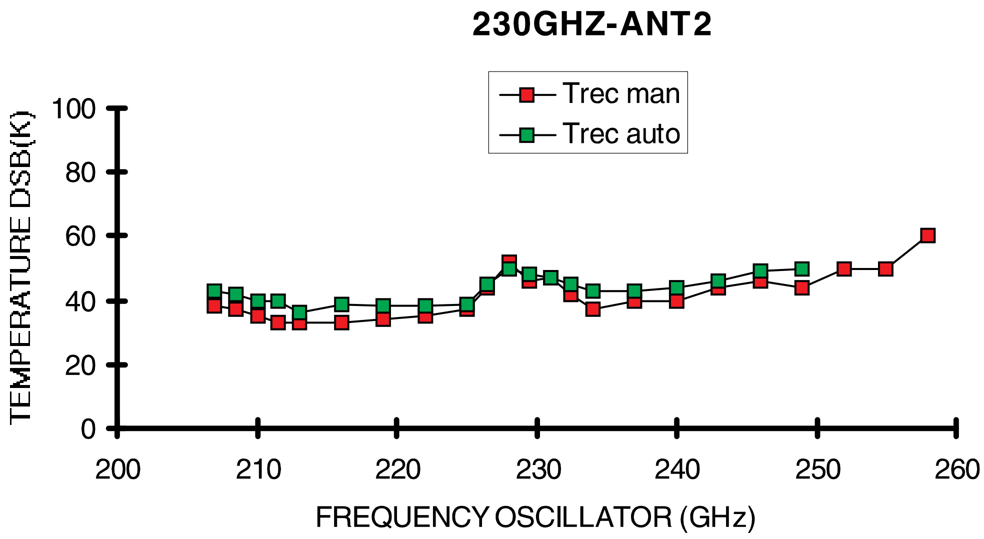
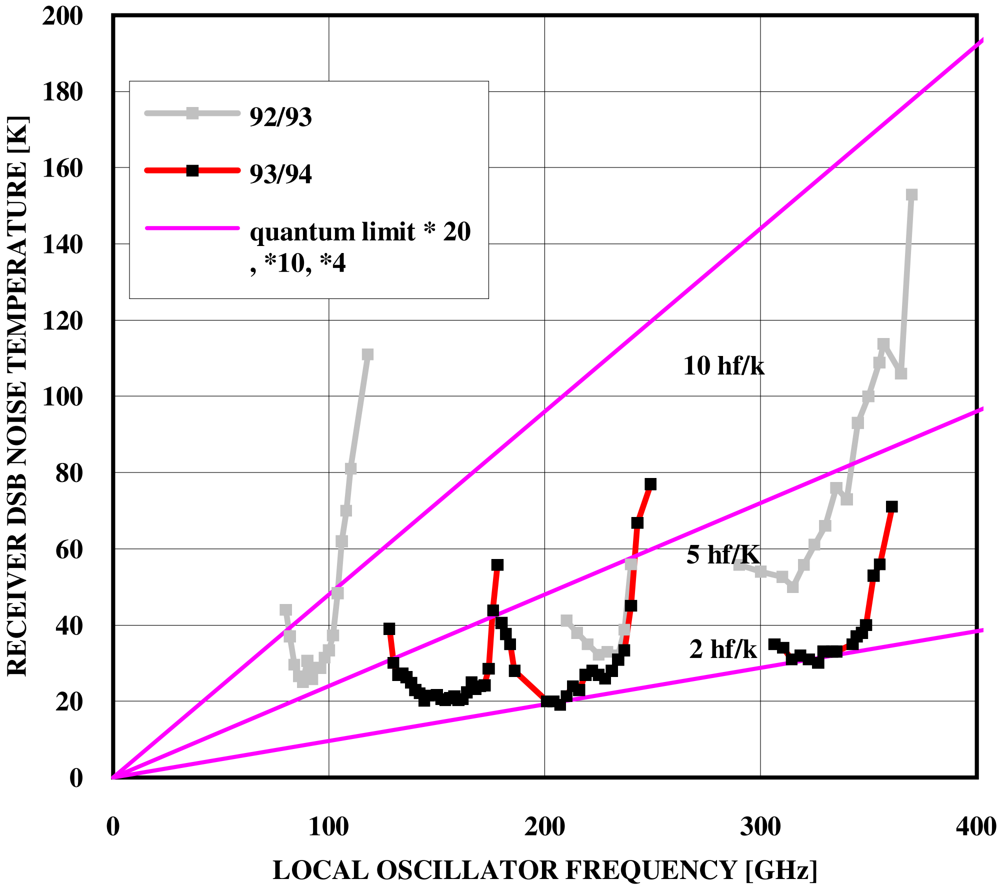


Figure 3.6: DSB noise performance of one of the 230 GHz receivers in operation at the PdB interferometer

Figure 3.7: DSB noise performance of the mm receivers constructed at IRAM over the frequency range 80 – 350 GHz. [Karpov 1999]



Chapter 4

Cross Correlators

Helmut Wiesemeyer

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

4.1 Introduction

As we already learned in the lecture on radio interferometry by S. Guilloteau (Chapter 2), the interferometer measures the complex cross-correlation function of the voltage at the outputs of a pair of antennas (i, j). This quantity, $R_{\rm ij}(\tau)$ is defined as

$$R_{\rm ij}(\tau) = \langle v_i(t) v_j(t+\tau) \rangle \tag{4.1}$$

(the brackets indicate the time average, see Appendix A). The cross-correlation function is related to the visibility function $V = |V| \exp{(i\varphi_{\rm SKY})}$ by

$$R_{\rm ij} = A_0 |V| \Delta\nu_{\rm IF} \cos{(2\pi\nu_{\rm IF}\tau_{\rm G} - \varphi_{\rm SKY})} \tag{4.2}$$

where A_0 is the collecting area of the antenna. Eq. 4.2 only holds for a quasi-monochromatic signal, $\Delta\nu_{\rm IF} \ll \nu_{\rm IF}$ (.e. the bandpass may be represented by a δ-function). The signal phase varies with time due to source structure and atmospheric perturbations (expressed by $\varphi_{\rm SKY}$), and due to the geometric delay $\tau_{\rm G}$. The timescale that is needed to fully sample a spectral line, given by the sampling theorem (see below) is much shorter. Here are examples of the different timescales:

1. timescale for phase variation by 1° due to source structure (for a point source at 98 GHz with $\Delta\alpha = 10''$ offset from phase reference center, east-west baseline of 258 m during transit): **2 min**

2. timescale for phase variation due to atmospheric perturbations: (depending on atmospheric conditions and baseline length): **1 sec − several hours**

3. sampling time step for a 80 MHz bandwidth: $\Delta\tau = 6.25$ns

4. maximum time lag needed for a 40 kHz resolution $\Delta\tau = 25\,\mu$s

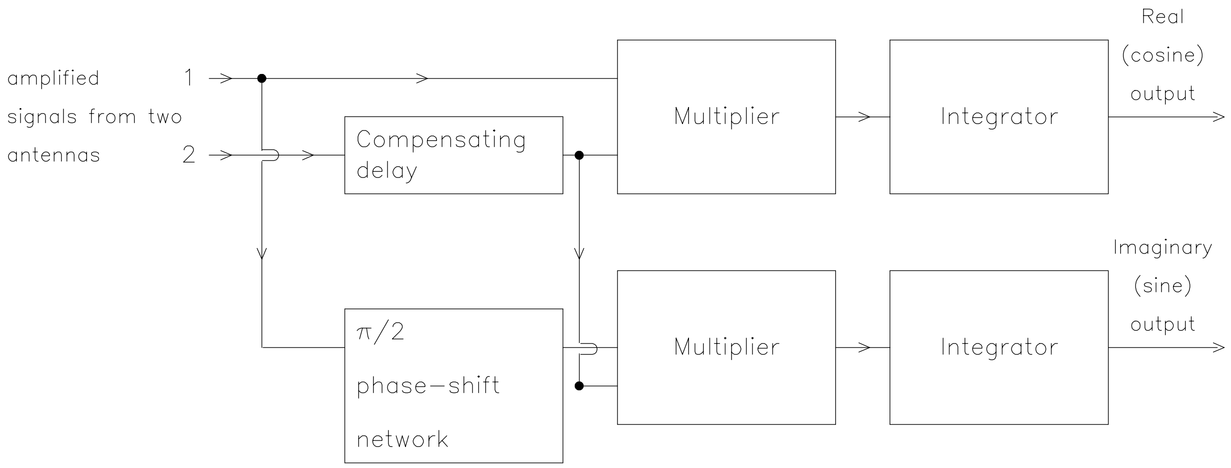


Figure 4.1: Architecture of a complex continuum cross correlator.

In the following, I will discuss digital techniques to evaluate $R_{\rm ij}(\tau)$. Analog methods of signal processing are highly impractical in radio interferometry, for mainly two reasons:

1. In time domain, high precision is needed.

2. The signal needs to be identically copied, in order to cross-correlate the output of one antenna with the outputs from all other antennas. This can be more easily done with digital techniques, than with analog ones.

The first signal processing steps are analog, beginning with the mixing in the heterodyne receivers. For reasons that will become clear later (see R. Lucas, Chapter 5), only the case of single-sideband reception is considered. The sidebands may be separated by a periodic phase shift of $\pi/2$ applied to the local oscillator. The signals are demodulated in two different ways by the correlator. At the entry of the correlator, filters are inserted, that are used to select the intermediate frequency bandpass. The following signal processing steps are digitally implemented, and are performed within the correlator:

1. Sampling the signal: in order to digitize the signal, it needs to be sampled. Bandwidth-limited signals (i.e. containing frequencies between zero and $\Delta\nu$) may be sampled without loss of interformation if the samples are taken at time intervals $\Delta t \leq 1/(2\Delta\nu)$.

2. In order to numerically compute the cross correlation function, the signals have to be discretized. The data are affected by such a quantization, but may be corrected for it. However, the loss of information cannot be recovered and degrades the correlator sensitivity.

3. Delay compensation: the geometric delays are eliminated for signals received from the direction of the pointing center. Remaining delays are due to source structure.

4. Until now, everything is done in the time domain. However, for spectroscopic applications, the desired output is the cross power spectral density, and not the cross correlation function. These quantities are Fourier-transform pairs (*Wiener-Khintchine* theorem). The transformation can be efficiently done by a processor performing a *Fast Fourier Transform*.

The plan of this lecture is as follows: after the basic theory, I will talk about the correlator in practice. Both intrinsic limitations, and system-dependent performance will be discussed. For further reading, the book of [Thompson et al 1986] (chapters 6 – 8), and the introduction by [D'Addario 1989] are recommended. Finally, as an example, the current correlator system on Plateau de Bure, and its upgrade in the near future, will be presented.

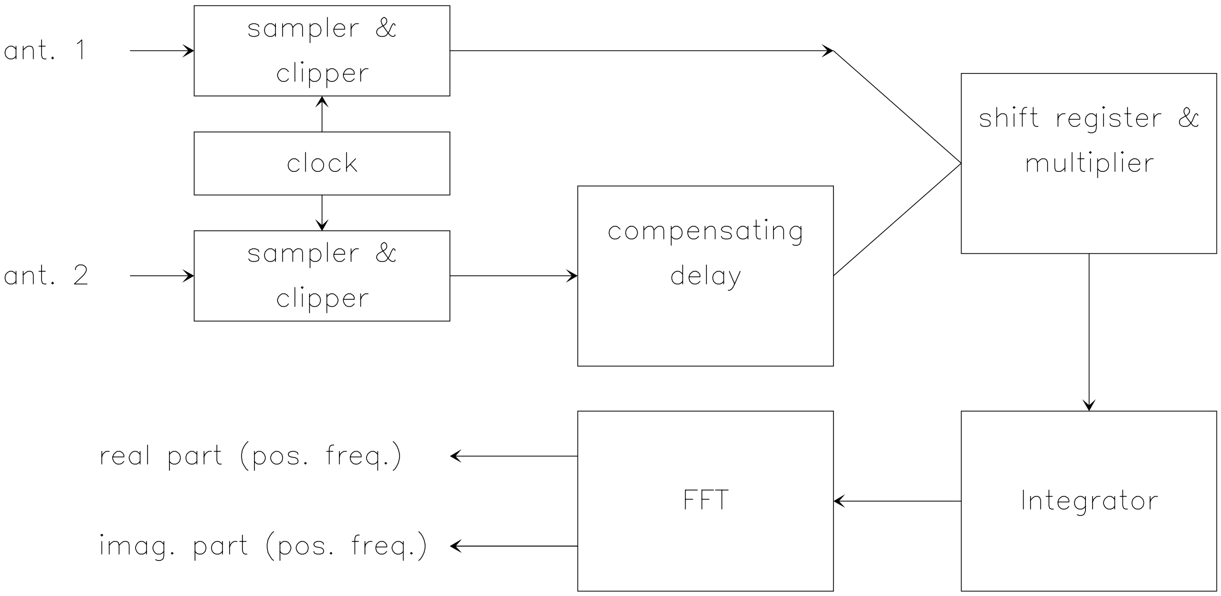


Figure 4.2: Architecture of a complex spectroscopic cross correlator.

4.2 Basic Theory

The "heart" of a correlator consists of the sampler and the cross-correlator. Eq. 4.2 represents an over-simplified case, because the bandwidth of the signals is neglected. The correlator output is rather modified by the Fourier transform of the bandpass function. For the sake of simplicity, let us assume an idealized rectangular passband of width $\Delta\nu_{\rm IF}$ for both antennas, centered at the intermediate frequency $\nu_{\rm IF}$, i.e.

$$|H_{\rm i}(\nu)| = |H_{\rm j}(\nu)| = \begin{cases} H_0, & |\nu - \nu_{\rm IF}| < \Delta\nu_{\rm IF}/2 \\ 0, & |\nu - \nu_{\rm IF}| > \Delta\nu_{\rm IF}/2 \end{cases} \tag{4.3}$$

(this assumption will be relaxed later). The correlator response to this bandpass is the Fourier transform of the cross power spectrum $H_{\rm i}(\nu)H_{\rm j}(\nu)^*$, which is shown in Fig. 4.3:

$$\int_0^\infty H_{\rm i}(\nu_{\rm IF})H_{\rm j}^*(\nu_{\rm IF})\exp{(2\pi i\nu_{\rm IF}\tau)}d\nu_{\rm IF} = H_0^2\Delta\nu_{\rm IF}\frac{\sin{(\pi\Delta\nu_{\rm IF}\tau)}}{\pi\Delta\nu_{\rm IF}\tau}\exp{(i2\pi\nu_{\rm IF}\tau)} \tag{4.4}$$

The correlator output consists of an oscillating part, and a $\sin(x)/x$ envelope. If the delay τ becomes too large, the sensitivity will be significantly decreased due to the sinc function (see Fig. 4.3). Strictly speaking, this is the response to the real part of the bandpass, which is symmetric with respect to negative frequencies. The imaginary part of the bandpass is antisymmetric with respect to negative frequencies, thus the correlator response is different. The separation of real and imaginary parts in continuum and spectroscopic correlators will be discussed below.

This example shows that accurate delay tracking (fringe stopping) is needed, if the bandwidth is not anymore negligible with respect to the intermediate frequency. In other words, the compensating delay $\tau_{\rm i}$ needs to keep the delay tracking error $\Delta\tau = \tau_{\rm G} - \tau_{\rm i}$ at a minimum. The offset $k\Delta t$ introduced in correlator channel k needs to be applied with respect to a fixed delay. In the following, the correlator response to a rectangular bandpass will be expressed by the more general instrumental gain function $G_{\rm ij}(\tau)$, defined by

$$A_0\int_0^\infty H_{\rm i}(\nu)H_{\rm j}^*(\nu)\exp{(2\pi i\nu\tau)}d\nu) = G_{\rm ij}(\tau)\exp{(2\pi i\nu_{\rm IF}\tau)} \tag{4.5}$$

$G_{\rm ij}(\tau) = |G_{\rm ij}(\tau)|\exp{(i\Phi_{\rm G})})$ is a complex quantity, including phase shifts due to the analog part of the receiving system (amplificators, filters)[1]. After fringe stopping, the single-sideband response of correlator

[1]Because τ is restricted to a maximum time lag, this instrumental gain factor does not describe long-term variations.

channel k becomes (for details, see R. Lucas, Chapter 5)

$$\begin{aligned} R_{\mathrm{ij,k}}() = R_{\mathrm{ij}}(k\Delta t) &= |V||G_{\mathrm{ij}}|\mathrm{Re}\left\{\exp\left(\pm 2\pi i\nu_{\mathrm{IF}}(\tau + k\Delta t) - i\varphi_{\mathrm{SKY}} \pm i\varphi_{\mathrm{G}}\right)\right\} \\ &= |V||G_{\mathrm{ij}}|\cos\left(\pm 2\pi\nu_{\mathrm{IF}}(\tau + k\Delta t) - \varphi_{\mathrm{SKY}} \pm \varphi_{\mathrm{G}}\right) \end{aligned} \tag{4.6}$$

where the plus sign refers to upper sideband reception, and the minus sign refers to lower sideband reception. From Eq. 4.6, we immediately see that the residual delay error (due to a non-perfect delay tracking) enters as a constant phase slope across the bandpass (with opposed signs in the upper and lower sidebands). The effect of such a phase slope on sensitivity will be discussed later. In order to determine the phase of the signal, the imaginary part of $R_{\mathrm{ij}}(\tau)$ has to be simultaneously measured. In a continuum correlator (Fig. 4.1), a $\pi/2$ phase shift applied to the analog signal yields the imaginary part. The signals are then separately processed by a cosine and a sine correlator [2]. In other words: the pattern shown in Fig. 4.3 is measured in the close vicinity of two points, namely at the origin, and at a quarter wave later, i.e. at $\tau = 1/(4\nu_{\mathrm{IF}})$. Note, however, that due to the sinc-envelope, the decreasing response function cannot be neglected if the bandwidth is comparable to the intermediate frequency.

In a spectroscopic correlator (Fig. 4.2), the imaginary part can be entirely deduced from the digitized signal: if N_{ch} is the number of complex spectral channels, $2N_{\mathrm{ch}}$ time lags are used, covering delays from $-N_{\mathrm{ch}}\Delta t$ to $(N_{\mathrm{ch}}-1)\Delta t$. The correlator output is a real signal with even and odd components (with respect to time lags of opposed signs). The N complex channels of the Fourier transform at positive frequencies yields the cross-power spectrum:

$$r_{\mathrm{ij,k}}(\nu_{\mathrm{IF}}) = r_{\mathrm{ij}}(k\delta\nu_{\mathrm{IF}}) = \int_{-\infty}^{\infty} R_{\mathrm{ij}}(t)\exp\left(2\pi i\nu_{\mathrm{IF}}t\right)dt \tag{4.7}$$

$$= \sum_{l=-N_{\mathrm{ch}}}^{N_{\mathrm{ch}}-1} \langle v_{\mathrm{i}}(t)v_{\mathrm{j}}(t+\tau+l\Delta t)\rangle \exp\left(2\pi ilk/2N_{\mathrm{ch}}\right) \tag{4.8}$$

(for channel k of a total of N_{ch} complex channels). The last expression represents the discrete Fourier transform. According to the symmetry properties of Fourier transforms, the even component of the correlator output becomes the real part of the complex spectrum, and the odd component becomes the imaginary part. The Fourier transform is efficiently evaluated using the Fast-Fourier algorithm. In practice, it is rather the digital measurement of the cross-correlation function $R_{IF}(\tau)$ that is non-trivial. It will be discussed in detail in Section 3. The ensemble of cross-power spectra $r_{ij}(\nu_{\mathrm{IF}})$, after tracking the source for some time, becomes (after calibration and several imaging processes) a channel map.

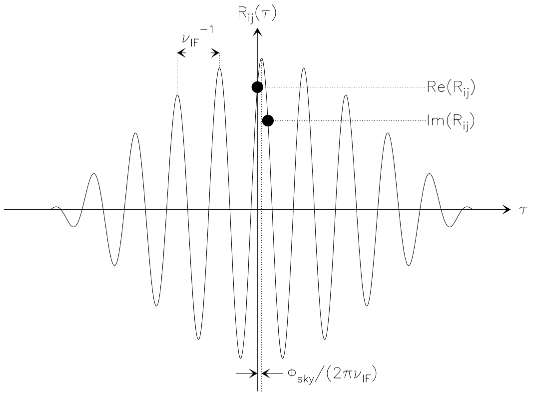
4.3 The Correlator in Practice

In order to numerically evaluate the cross-correlation function R_{ij}, the continuous signals entering the cross correlator need to be sampled and quantized. According to Shannon's sampling theorem [Shannon 1949], a bandwidth-limited signal may be entirely recovered by sampling it at time intervals $\Delta t \leq 1/(2\Delta\nu_{\mathrm{IF}})$ (also called sampling at Nyquist rate). The discrete Fourier transform of the sufficiently sampled cross-correlation function theoretically yields the cross-power spectrum without loss of information. However, in practice, two intrinsic limitations exist:

- In order to discretize a signal, it is not only sampled, it also has to be quantized. The cross-correlation function, as derived from quantized signals, does not equal the cross-correlation function of continuous signals. Moreover, the sampling theorem does not hold anymore for quantized signals. The reasons will become clear below.

- Eq. 4.7 theoretically extends from $-\infty$ to $+\infty$. In practice (Eq. 4.8), only a maximum time lag can be considered: limited storage capacities and digital processing speed are evident reasons, another limiting factor are the different timescales mentioned before. The abrupt cutoff of the time window affects the data.

[2] For the sake of completeness, it should be mentioned that this is a special case of the so-called *Hilbert* transform, which property is to change signal phases by $\pi/2$, but to leave amplitudes unchanged.



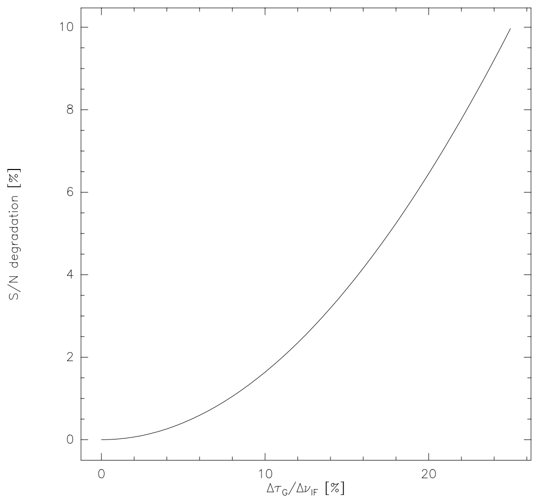


Figure 4.3: Left: Correlator output (single-sideband reception) for a rectangular passband with $\Delta\nu_{\rm IF}/\nu_{\rm IF} = 0.2$. Due to the signal phase $\varphi_{\rm SKY}$, the oscillations moves through the sinc envelope by $\varphi_{\rm SKY}/2\pi\nu_{\rm IF}$. The shift may also be due to the phase of the complex gain (in this case, the shift would be in opposed sense for USB and LSB reception). Right: Sensitivity degradation due to a delay error $\Delta\tau_{\rm G}$ (with respect to the inverse IF bandwidth). The effect is due to the fall-off of the sinc envelope.

These "intrinsic" limitations are discussed in Sections 4.3.1 and 4.3.2. The system-dependent performance will be addressed in Section 4.3.3.

4.3.1 Digitization of the input signal and clipping correction

As already mentioned, sampling at the Nyquist rate retains all information. However, quantizing the input signal leads to a loss of information. This can be qualitatively understood in the following way: in order to reach the next discrete level of the transfer function, some offset has to be added to the signal. If the input signal is random noise of zero mean, the offset to be added will also be a random signal of zero mean. In other words, a "quantization" noise is added to the signal, that leads to a loss of information. In addition, the added noise is not anymore bandwidth limited, and the sampling theorem does not apply: oversampling will lead to improved sensitivity.

Many quantization schemes exist (see e.g. [Cooper 1970]). It is entirely sufficient to use merely a few quantum steps, if the cross-correlation function will be later corrected for the effects of quantization. For the sake of illustration, the transfer function of a four-level 2-bit quantization is shown in Fig. 4.5. Each of the four steps is assigned a sign bit, and a magnitude bit. After discretizing the signal, the samples from one antenna are shifted in time, in order to compensate the geometric delay $\tau_{\rm G}(t)$. The correlator now proceeds in the following way: for each delay step Δt, the corresponding sign and magnitude bits are put into two registers (one for the first antenna, and one for the second). The second register is successively shifted by one sample. In this way, sample pairs from both antennas, separated by a successively longer time lag, are created. These pairs are multiplied, using a multiplication table. For the case of four-level quantization, it is shown in Fig. 4.5. Products which are assigned a value of $\pm n^2$ are called "high-level products", those with a value of $\pm n$ are "intermediate-level products", and those with a value of ± 1 "low-level products". The products (evaluated using the multiplication table in Fig. 4.5) are send to a counter (one counter for each channel, i.e. for each of the discrete time lags). After the end of the integration cycle, the counters are read out.

In practice, the multiplication table will be shifted by a positive offset of n^2, to avoid negative products (the offset needs to be corrected when the counters are read out). This is because the counter is simply an adding device. As another simplification, low-level products may be deleted. This makes digital implementation easier, and accounts for a loss of sensitivity of merely 1 % (see Table 4.1). Finally, not all bits of the counters' content need to be transmitted (see Section 3.3.2).

Before the normalized contents of the counters are Fourier-transformed, they need to be corrected, because the cross-correlation function of quantized data does not equal the cross-correlation function of continuous data. This "clipping correction" can be derived using two different methods. As an example for the case of full 4-level quantization:

- Four-level cross-correlation coefficient according to the multiplication table Fig. 4.5. The cross-correlation coefficient ρ is a normalized form of the cross-correlation function (see Appendix A):

$$\rho_4 = \frac{2n^2(N_{01,01} - 2N_{01,11}) + 4n(N_{00,01} - N_{00,11}) + 2(N_{00,00} - N_{00,10})}{2(n^2 N_{01,01} + N_{00,00})_{\rho=1}} \tag{4.9}$$

  where $N_{\mathrm{ij,kl}}$ is the number of counts with sign bit i and magnitude bit j at time t (first antenna), and sign bit k and magnitude bit l at time $t+\tau$ (second antenna). $\pm n$ is the product value assigned to intermediate-level products.

- Clipping correction, first method: evaluate the $N_{\mathrm{ij,kl}}$ in Eq. 4.9, using joint probabilities $P_{\mathrm{ij,kl}}$ (see Appendix A for the definition of the jointly Gaussian probability distribution), such as

$$N_{01,01} = NP_{01,01} = \frac{N}{2\pi\sigma^2\sqrt{1-\rho^2}} \int_{v_0}^{\infty}\int_{v_0}^{\infty} \exp\left[\frac{-(x^2+y^2-2\rho xy)}{2\sigma^2(1-\rho^2)}\right] dx dy \tag{4.10}$$

  (N is the number of signal pairs, separated by the time lag of the channel under consideration, v_0 is the clipping voltage, see Fig. 4.4).

- Clipping correction, second method: using Price's theorem for functions of jointly random variables. The result, derived in Appendix B, is shown in Fig. 4.4:

$$\begin{aligned} R_4 = \; & \tfrac{\sigma^2}{\pi} \int_0^\rho \tfrac{1}{\sqrt{1-r^2}} \Bigg\{ (n-1)^2 \left[\exp\left(\tfrac{-v_0^2}{\sigma^2(1+r)}\right) + \exp\left(\tfrac{-v_0^2}{\sigma^2(1-r)}\right)\right] \\ & + 4(n-1)\exp\left(\tfrac{-v_0^2}{2\sigma^2(1-r^2)}\right) + 2 \Bigg\} dr\,. \end{aligned} \tag{4.11}$$

Although the discrete, normalized cross-correlation function and the continuous cross-correlation coefficient are almost linearly dependent within a wide range, the correction is not trivial. An analytical solution is only possible for the case of two-level quantization ("van Vleck correction" [Van Vleck 1966]).

In practice, several methods are used to numerically implement Eq. 4.11 (in the following, the index k means k-level quantization). The integrand may be replaced by an interpolating polynomial, allowing to solve the integral. One may also construct an interpolating surface $\rho(R_{\mathrm{k}}, \sigma)$. As already discussed, the clipping correction cannot recover the loss of sensitivity due to quantization. The loss of sensitivity for k-level discretization may be found by evaluating the signal-to-noise ratio

$$\Re_{\mathrm{sn},k} = \frac{R_{\mathrm{k}}}{\sigma_k} = \frac{R_{\mathrm{k}}}{\sqrt{\langle R_k^2\rangle - \langle R_k\rangle^2}} \tag{4.12}$$

In order to minimize the loss of sensitivity, the clipping voltage (with respect to the noise σ) needs to be adjusted such that the correlator efficiency curve in Fig. 4.4 is at its maximum. The correlator efficiency is defined with respect to the signal-to-noise ratio of a (fictive) continuous correlator, i.e.

$$\eta_k = \frac{\Re_{\mathrm{sn},k}}{\Re_{\mathrm{sn},\infty}} = \frac{\Re_{\mathrm{sn},k}}{\rho\sqrt{N_{\mathrm{q}}}} \tag{4.13}$$

where N_{q} is the number of samples. Table 4.1 summarizes the results for different correlator types and samplings.

Due to the discretization of the input voltages (as shown in Fig. 4.5), any knowledge of the absolute signal value is lost. The signal amplitude is recovered by a regularly performed calibration (using a calibration load of known temperature, for details, see the lecture by A. Dutrey).

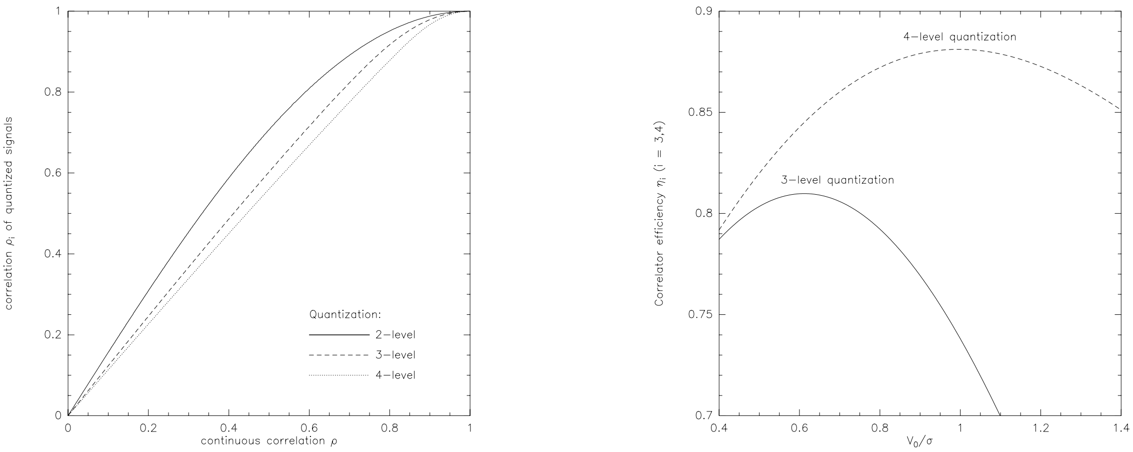
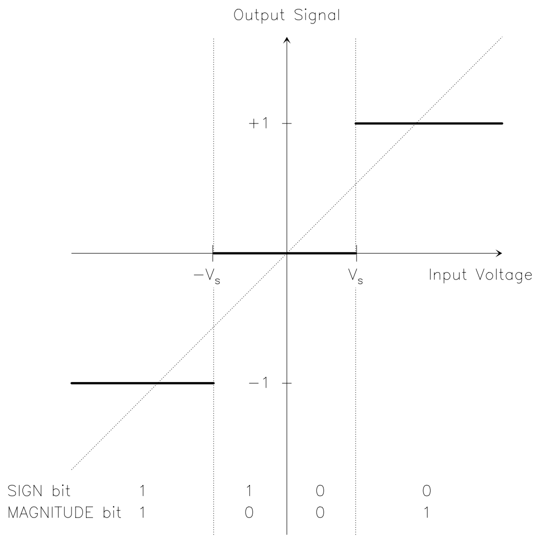


Figure 4.4: Left: Clipping correction (cross correlation coefficient of a continuous signal vs. cross correlation correlation coefficient of a quantized signal) for two-, three- and four level quantization (with optimized threshold voltage). The case of two level quantization is also known as *van Vleck correction*. For more quantization levels, the clipping correction becomes smaller. Right: Correlator efficiency as function of the clipping voltage, for three-level and four-level quantization (at Nyquist sampling).



	S(x)	0	0	1	1
	M(x)	1	0	0	1
S(y)	M(y)				
0	1	n^2	n	$-n$	$-n^2$
0	0	n	1	−1	$-n$
1	0	$-n$	−1	1	n
1	1	$-n^2$	$-n$	n	n^2

Figure 4.5: Left: Transfer function for a 4-level 2-bit correlator. The dashed line corresponds to the transfer function of a (fictive) continuous correlator with an infinite number of infinitesimally small delay steps. Right: Multiplication table. $S(x)$ is the signal bit at time t, $M(x)$ is the magnitude bit at time t (respectively $S(y)$ and $M(y)$ at time $t+\tau$).

Table 4.1: Correlator parameters for several quantization schemes

method	n	v_0 $[\sigma_{\rm rms}]$	$\eta_{\rm q}^{(1)}$ for sampling rate	
			$2\Delta\nu_{\rm IF}^{(2)}$	$4\Delta\nu_{\rm IF}^{(3)}$
two-level	–	–	0.64	0.74
three-level	–	0.61	0.81	0.89
four-level	3	1.00	$0.88^{(4)}$	0.94
	4	0.95	0.88	0.94
∞-level	–	–	1.00	1.00

Notes:
(1) The correlator efficiency is defined by Eq. 4.13.
The values are for an idealized (rectangular)
bandpass and after level optimization.
(2) Nyquist sampling,
(3) oversampling by factor 2
(4) 0.87 if low level products deleted
(case of Plateau de Bure correlator)

Table 4.2: Time lag windows

Description	Lag window	Spectral window
rectangular	$w(t) = 1$ for $\lvert t\rvert \le \tau_{\rm m}$, else 0	$\hat{w}(\nu) = 2\tau_{\rm m}\frac{\sin(2\pi\nu\tau_{\rm m})}{2\pi\nu\tau_{\rm m}}$
Bartlett	$w(t) = 1 - \frac{\lvert t\rvert}{\tau_{\rm m}}$ for $\lvert t\rvert \le \tau_{\rm m}$, else 0	$\hat{w}(\nu) = \tau_{\rm m}\left(\frac{\sin(\pi\nu\tau_{\rm m})}{\pi\nu\tau_{\rm m}}\right)^2$
von Hann	$w(t) = \frac{1}{2}\left(1+\cos(\frac{\pi t}{\tau_{\rm m}})\right)$ for $\lvert t\rvert \le \tau_{\rm m}$, else 0	$\hat{w}(\nu) = \tau_{\rm m}\cdot\frac{\sin(2\pi\nu\tau_{\rm m})}{2\pi\nu\tau_{\rm m}}\cdot\frac{1}{1-(2\nu\tau_{\rm m})^2}$
Welch	$w(t) = \left(1-(\frac{t}{\tau_{\rm m}})^2\right)$	$\hat{w}(\nu) = \frac{1}{(\pi\nu)^2\tau_{\rm m}}\left(\frac{\sin(2\pi\nu\tau_{\rm m})}{2\pi\nu\tau_{\rm m}} - \cos(2\pi\nu\tau_{\rm m})\right)$
Parzen	$w(t) = \begin{cases} 1-6\left(\frac{t}{\tau_{\rm m}}\right)^2+6\left(\frac{\lvert t\rvert}{\tau_{\rm m}}\right)^3 \\ \quad \text{for} \quad \lvert t\rvert \le \tau_{\rm m}/2 \\ 2\left(1-\frac{\lvert t\rvert}{\tau_{\rm m}}\right)^3 \\ \quad \text{for} \quad \tau_{\rm m}/2 < \lvert t\rvert \le \tau_{\rm m} \end{cases}$	$\hat{w}(\nu) = \frac{3}{4}\tau_{\rm m}\left(\frac{\sin(\pi\nu\tau_{\rm m}/2)}{\pi\nu\tau_{\rm m}/2}\right)^4$

4.3.2 Time lag windows and spectral resolution

According to the sampling theorem, we can recover the cross-power spectral density within a bandwidth $\Delta\nu_{\rm IF} = 1/(2\Delta t)$, if the sampling step is Δt. The channel spacing $\delta\nu$ is then determined by the maximum time lag $\tau_{\rm max} = N_{\rm ch}\Delta t$ (where $N_{\rm ch}$ is the number of channels), i.e.

$$\delta\nu = \frac{1}{2\tau_{\rm max}} = \frac{1}{2N_{\rm ch}\Delta t} \tag{4.14}$$

However, the data acquisition is abruptly stopped after the maximum time lag. After the Fourier transform, the observed cross power spectrum is convolved with the Fourier transform $\hat{w}(\nu)$ of the box-shaped time window $w(t)$, producing strong sidelobes:

$$w(\tau) = \left\{ \begin{array}{ll} 1, & |\tau| \leq \tau_{\rm max} \\ 0, & \text{else} \end{array} \right\} \rightleftharpoons \hat{w}(\nu) = 2\tau_{\rm max}\frac{\sin\left(2\nu\tau_{\rm max}\right)}{2\nu\tau_{\rm max}} \tag{4.15}$$

These oscillations are especially annoying, if strong lines are observed. They may be minimized, if the box-shaped time lag window is replaced by a function that rises from zero to peak at negative time lags, and decreases to zero at positive time lags (apodization). Such a window function suppresses the sidelobes, at the cost of spectral resolution. A comparison between several window functions is given in Fig. 4.6, together with sidelobe levels and spectral resolutions (defined by the FWHP of the main lobe of the spectral window). Table 4.2 summarizes the various functions in time and spectral domains. The default of the Plateau de Bure correlator is the Welch window, because it still offers a good spectral resolution. Moreover, the oscillating sidelobes partly cancel out the contamination of a channel by the signals in adjacent channels. Of course, the observer is free to deconvolve the spectra from this default window, and to use another time lag window.

Note: If you apodize your data, not only the effective spectral resolution is changed. Due to the suppression of noise at large time lags, the sensitivity is increased. The variance ratio of apodized data to unapodized data,

$$\int_{-\infty}^{\infty} |w(t)|^2 dt = \int_{-\infty}^{\infty} |\hat{w}(\nu)|^2 d\nu = 1/B_{\rm n} \tag{4.16}$$

defines the noise equivalent bandwidth $B_{\rm n}$. It is the width of an ideal rectangular spectral window (i.e. $\hat{w}(\nu) = 1/B_{\rm n}$ with zero loss inside $|\nu| \leq B_{\rm n}/2$, and infinite loss outside) containing the same noise power as the actual data. For sensitivity estimates of spectral line observations, the channel width to be used is thus the noise equivalent width, and neither the channel spacing, nor the effective spectral resolution. For commonly used time windows, Fig. 4.6 gives the noise equivalent bandwidths.

4.3.3 Main limitations

In real life, cross-correlators are subject to the performance of the whole receiving system. This comprises the "analog part" (the signal path from the receivers to the IF filters at the correlator entry), and the "digital part" (everything behind the sampler). Although the analog part is out of the correlator, its performance requires to change our assumptions concerning the input data. This complicates the analysis of the correlator response. The following discussion refers to instantaneous errors only. However, in interferometric mapping, scan-averaged visibilities are used, and the data may be less affected.

Analog part

The shape of the bandpass function (amplitude and phase) at the correlator output is mainly due to the correlator's response to the filters inserted in the IF band at the correlator entry. So far, for the sake of simplicity, rectangular passbands, centered at the intermediate frequency $\nu_{\rm IF}$, have been assumed. A more complex (and more realistic) case may be an amplitude slope where the logarithm of the amplitude varies linearly with frequency. Although the bandpass function will be calibrated (see Eq. 4.17, and R. Lucas chapter 5), the effect of such a slope on sensitivity remains. A derivation of the signal-to-noise ratio for that case is beyond the scope of this lecture. To give an impression of the order of magnitude: a slope of

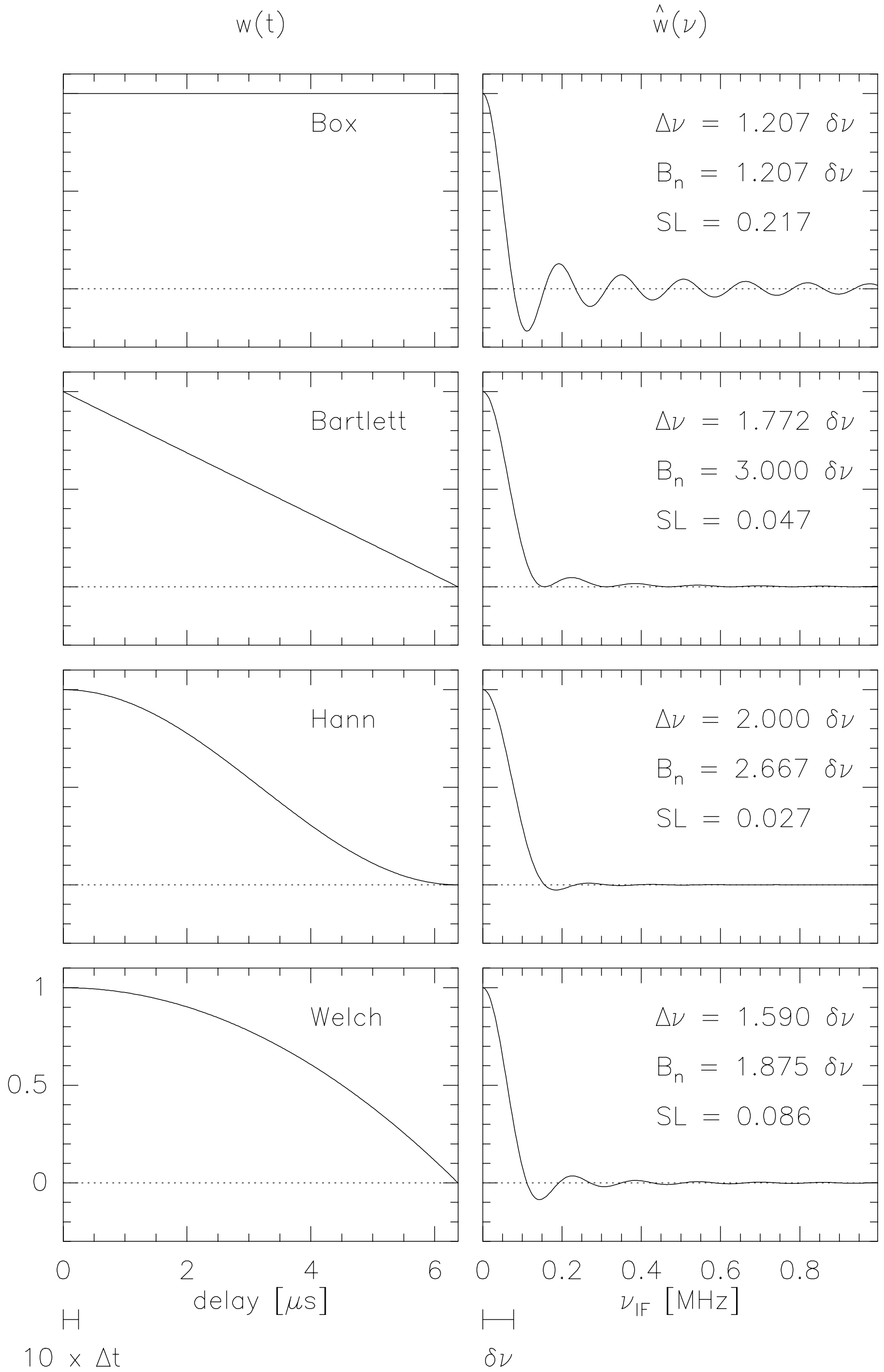


Figure 4.6: Several time lag windows, and their Fourier transforms (normalized to peak). The sidelobe levels SL are indicated, as well as the spectral resolution (defined as the FWHP of the main lobe), and the noise equivalent width. The delay stepsize, and channel spacing are indicated for the following example: 256 channels, clock rate 40 MHz, resulting in a channel spacing of 78.125 kHz.

Table 4.3: Effects of delay pattern on the sensitivity

Intermediate frequency bandwidth	$\nu_{\rm IF} = 160\,{\rm MHz}$
Baseline	$b = 100\,{\rm m}$
Zenith distance of source in direction $\mathbf{s}$	$\Theta = 30°$
Results in geometric delay:	$\tau_{\rm G} = \mathbf{b}\cdot\mathbf{s}/c = 0.17\,\mu{\rm s}$
Attenuation according to Eq. 4.4	$1\,\%$

3.5 dB (edge-to-edge) leads to a 2.5 % degradation of the sensitivity calculated for a rectangular passband. A center frequency displacement of 5 % of the bandwidth leads to the same degradation.

As already demonstrated, delay-setting errors linearly increase with the intermediate frequency (Eq. 4.6). Table 4.3 gives an impression of the decrease of sensitivity due to a delay error. For a range of delay errors $\Delta\tau$, the effect is also shown in Fig. 4.3. For example, a delay error of $0.12/\Delta\nu_{\rm IF}$ accounts for a 2.5 % degradation. Delay errors are mainly due to inaccurately known antenna positions (asking for better baseline calibration), or due to errors in the transmission cables.

Phase errors across the bandpass may also be of random nature. A phase fluctuation of 12.8° (rms) per scan leads to a degradation of $(1 - \exp(-\sigma_\Phi^2/2)) \times 100\,\% = 2.5\,\%$. Fluctuations across the bandpass also appear as ripples. They may have several reasons, and are mainly due to the *Gibbs phenomenon*, and due to reflections in the transmission cables. A sinusoidal bandpass ripple of 2.9 dB (peak-to-peak) yields a 2.5 % degradation in the signal-to-noise ratio. The *Gibbs phenomenon* also occurs in single-dish autocorrelation spectrometers. For the sake of illustration, let us again assume a perfectly flat response of receivers and filters. However, the filter response function is only be flat across the IF passband. Towards its boundary, steep edges occur. We already learned that strong spectral lines may show ripples, if no special data windowing in time domain is applied. The Gibbs phenomen is due to a similar problem (but now the spectral line is replaced by the edge of a flat rectangular band extending in frequency from zero to $\Delta\nu_{\rm IF}$). The output of the cosine correlator is symmetric, but the sine output (imaginary part) is antisymmetric, thus including an even steeper edge. Convolving this edge with the sinc function (i.e. the spectral window) results in strong oscillations. Let us call this function $f(\nu)$. For calibration purposes, the Gibbs phenomenon has to be avoided: the problem is that calibration uses the system response to a flat-spectrum continuum source. A source whose visibility is $V(\nu)$ is seen as $f(\nu) * (G_{\rm ij}(\nu)V(\nu))$ (where $G_{\rm ij}$ is now a frequency-dependent complex gain function). After calibration it becomes

$$\hat{V}(\nu) = \frac{f(\nu) * [G_{\rm ij}(\nu)V(\nu)]}{f(\nu) * G_{\rm ij}(\nu)} \tag{4.17}$$

You immediately see that the complex gain $G_{\rm ij}(\nu)$ does not cancel out, as desired, and $\hat{V}(\nu) \neq V(\nu)$. Automatic calibration procedures have to flag the channels concerned. As shown in Fig. 4.7, for the cosine correlator, the effect is strongest at the band edges, but the output of the sine correlator also shows ripples in the middle of the band (thus, the problem is of greater importance for interferometers than for single-dish telescopes using auto-correlators). If the bandwidth to be observed is synthesized by two adjacent frequency windows, the phenomenon is strongest at the band center anyhow. You should avoid to place your line there, if it is on top of an important continuum. To what extent the Plateau de Bure system is concerned, will be discussed in Section 4.4.1.

The above summary of the system-dependent performance of a correlator is not exhaustive. For example, the phase stability of tunable filters, which depends on their physical temperature, is not discussed. Alternatives to such filters are image rejection mixers (as applied in the Plateau de Bure correlator, see last section).

Digital part

Errors induced by the digital part are generally negligible with respect to the analog part. In digital delays, a basic limitation is given by the discrete nature of the delay compensation, which cannot be more accurate than given by the clock period of the sampler. However, digital techniques allow for high clock rates, keeping this error at a minimum.

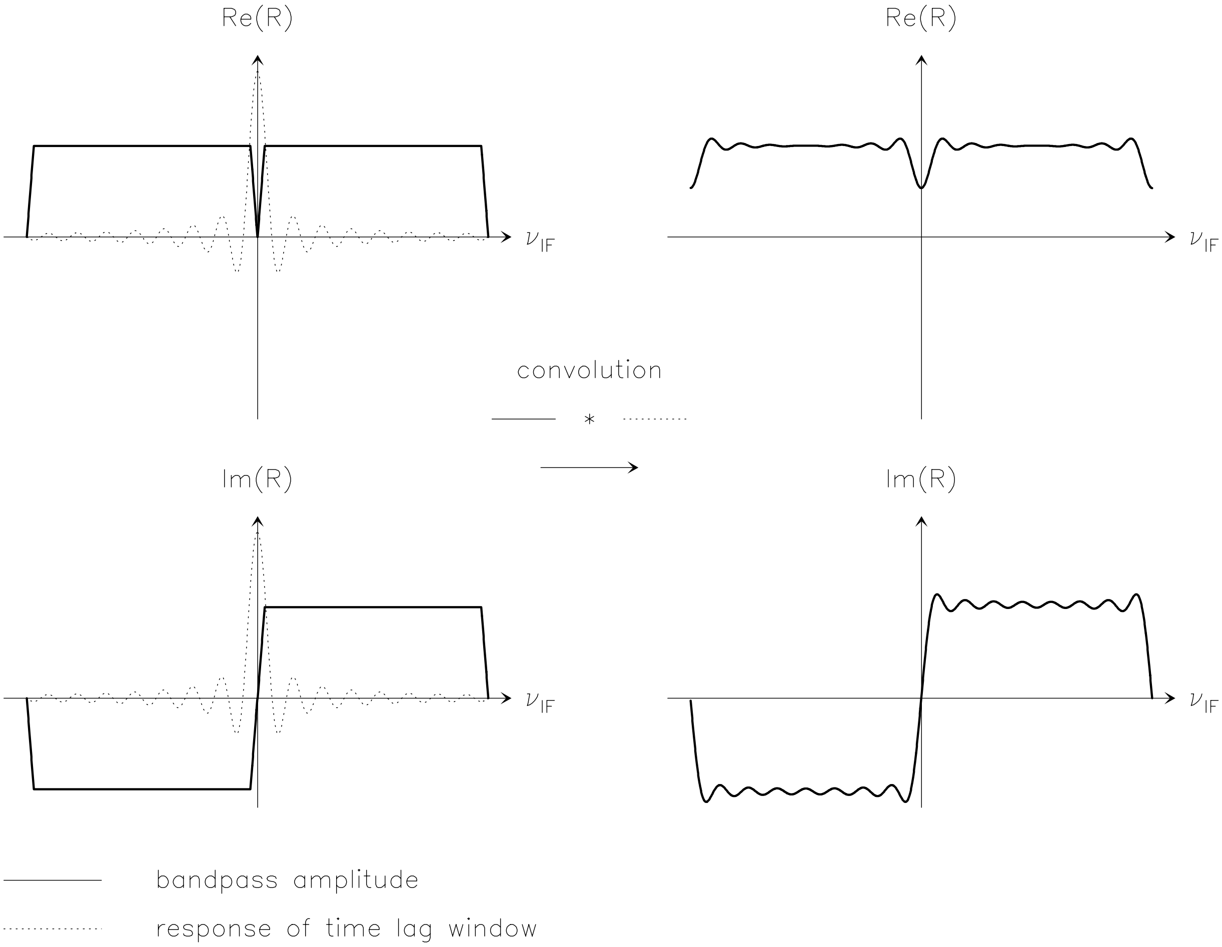


Figure 4.7: The Gibbs phenomenon. The convolution of the bandpass with the (unapodized) spectral window (sinc function) is shown for the real and imaginary parts. Note that for the real part, the phenomenon is strongest at the band edges, whereas for the imaginary part, it contaminates the whole bandpass.

Table 4.4: Maximum integration time of a 16-bit counter

clock frequency:	80 MHz
weight for intermediate-level products:	$n = 3$
positive offset:	$n^2 = 9$
weight for autocorrelation product:	18 (using offset multiplication table)
carry out rate of a 4-bit adder	$18/2^4 = 1.125$
maximum integration time:	$2^{16}/(80\,\mathrm{MHz} \times 1.125) = 0.73\,\mathrm{ms}$
same with a 4-bit prescaler:	$2^{16} \times 2^4/(80\,\mathrm{MHz} \times 1.125) = 11.7\,\mathrm{ms}$

Evidently, a basic limitation is given by the memory of the counters, setting the maximum time lag (which in turn defines the spectral resolution, as already discussed): with $2K$ bits, we can exactly represent N^{2K} numbers. However, the information contained in the bits is not equivalent. For the 3-level correlator, the output of each channel $i = 1, ..., N$ is

$$R(i) = \frac{1}{2}\left(N \pm \sqrt{N}\sqrt{1 - \mathrm{erf}\left(v_0/\sqrt{2}\right)}\right) \tag{4.18}$$

(assuming white, Gaussian noise of zero mean and of unit variance, and neglecting the weak contribution of the astrophysical signal). The 1σ-precision of the output is $\approx \sqrt{N}/2$, contained in the last $K-1$ bits, which thus do not need to be transmitted. The maximum integration time before overflow occurs is set by the number of bits of the counter, and the clock frequency. Table 4.4 shows an example.

The only error cause due to the correlator that is worth to be mentioned is the sampler, i.e. the analog-to-digital conversion. As already shown, the threshold levels are adjusted with respect to the noise in the unquantized signal. However, the noise power may change during the integration. In that case, the correlator does not operate anymore at its optimum level (see Fig. 4.4). This error cause can be eliminated with an automatic level control circuit. However, slight deviations from the optimal level adjustment may remain. Without going too far into detail, the deviations can be decomposed in an even and an odd part: in one case, the positive and negative threshold voltages move into opposed directions (even part of the threshold error). The resulting error can be equivalently interpreted as a change of the signal level with respect to the threshold v_0, and leads to a gain error. In the other case, the positive and negative threshold voltages move into the same direction (odd part of the threshold error). This error, however, can be reduced by periodic sign reversal of the digitized samples (if the local oscillator phase is simultaneously shifted by π, the correlator output remains unaffected). Combining the original and phase-shifted outputs, the error cancels out with high precision. Such a phase shift is implemented in the first local oscillators of the Plateau de Bure system (for details see the lecture by R. Lucas). Note also that threshold errors of up to 10 % can be tolerated without degrading the correlator sensitivity too much: the examination of Fig. 4.4 shows that such an error results in a signal-to-noise degradation of less than 0.2 % for a 3-level system, and of less than 0.5 % for a 4-level system (the maxima of the efficiency curves are rather broad).

Another problem is that the nominal and actual threshold values may differ. The error can be described by "indecision regions". By calculating the probability that one or both signals of the cross-correlation product fall into such an indecision region, the error can be estimated. With an indecision region of 10 % of the nominal threshold value, the error is negligibly small.

Finally, it should be noted that strict synchronisation between the cosine and the sine correlators is needed. Any deviation will introduce a phase error.

4.4 The correlator on Plateau de Bure

As an example of a cross-correlator used in mm-wave interferometry, I briefly introduce the correlator system on Plateau-de-Bure. Only a spectroscopic correlator is in use. Continnum bands are synthesized by channel averages covering the desired bandwidths. Aspects concerning concrete observing projects are addressed in R. Neri's lecture (Chapter 6).

Table 4.5: The Complex Cross Correlator on Plateau de Bure

Bandwidth [MHz]	Sideband of IRM$^{(1)}$	Clock Rate [MHz]	Time Multiplex Factor	Number of Lags (2)	Complex Channels	Channel Spacing [MHz]	Spectral Resolution [MHz] (3)	(4)
160	USB + LSB	40	4	64	64	2.500	3.025	4.000
80	USB + LSB	40	4	128	128	0.625	0.756	1.000
40	USB + LSB	40	2	256	256	0.156	0.189	0.250
20	USB	40	2	512	256	0.078	0.094	0.125
10	USB	20	1	512	256	0.039	0.047	0.062

Notes: (1) image rejection mixer (2) with negative & positive time lags (3) box-shaped time-lag window (4) Welch time-lag window

4.4.1 The current system

The cross-correlator currently in use on Plateau de Bure, called CAMEMBERT[3], comprises six independent units. It consists of three parts: an IF processor (frequency setting, low-pass filter selection, oscillator phase control – i.e. the analog functions), a digital part, controlled by a master processor (i.e. delay steps, clipping correction, FFT, small delay corrections, bandpass correction), and a satellite micro reading out and further processing the correlations.

Each unit can be placed in the [100, 600] MHz IF band, in steps of 0.625 MHz (by using a third frequency conversion). The 40, 80 and 160 MHz bandwidths are synthesized by the adjacent upper and lower sidebands of an image rejection mixer. Low-pass filters select the desired bandwidth. Note that these bandwidths show the Gibbs phenomenon right in the middle of the band (i.e. at the edges of the IRM sidebands). The central two channels are flagged by default, the observer should avoid to place the most important part of the line there. The 10 and 20 MHz bandwidths do not have this restriction: for the 20 MHz bandwidth, only the upper sideband of the IRM is used. The 10 MHz bandwidth (only unit 5) is produced by slowing down the clock rate from 40 to 20 MHz. The spectroscopic capabilities of the cross-correlator at Plateau de Bure are summarized in Table 4.5. Part of the flexibility is achieved by using the "time-multiplexing" technique. For example, a time-multiplexing factor four means that the data, arriving at a rate of 160×10^6 samples/s, are alternately put into four shift-registers. The shift registers are read out at the clock frequency of 40 MHz, thus creating four data streams taken at a rate that is lower by a factor of four (as compared to the sampling speed). Equivalently, a time-multiplex factor two means two data streams at a rate of 80 MHz each.

4.4.2 Future improvements

The next-generation correlator for the Plateau de Bure interferometer will allow for more flexibility, due to the following improvements:

- global bandwidth: 2.56 GHz (vs. 0.96 GHz in the current system),
- flexibility: 8 units with channel spacings in powers-of-2 sequence (vs. 6 units, channel spacing in powers-of-4 sequence)
- global digital performance: 9.8 Teramultiplications per second (vs. 1.3 TM s^{-1}).

These improvements are made possible by using new, more integrated technology at both analog and digital signal processing steps.

[3]an acronym for *Correlator Architecture for Multi-Element or Multi-BEam Radio-Telescopes*

4.5 Appendix

4.5.1 Summary of definitions

- Cross-correlation function of voltage outputs v_{i} and v_{j} from antenna pair (i, j):

$$R_{ij}(\tau) = \langle v_{\mathrm{i}}(t) v_{\mathrm{j}}(t+\tau)\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T v_{\mathrm{i}}(t) v_{\mathrm{j}}(t+\tau) dt \tag{4.19}$$

- Covariance of two jointly random variables:

$$\mu = \langle xy\rangle - \langle x\rangle\langle y\rangle \tag{4.20}$$

For signals of zero mean, and again identifying $x = v_{\mathrm{i}}(t)$ and $y = v_{\mathrm{j}}(t+\tau)$,

$$\mu = R_{ij}(\tau) \tag{4.21}$$

- Cross-correlation coefficient of two jointly random variables x, y of variance σ_{x}^2 and σ_{y}^2:

$$\rho = \frac{\mu}{\sigma_{\mathrm{x}}^2 \sigma_{\mathrm{y}}^2} \tag{4.22}$$

For jointly normal random variables of zero mean and of variance $\sigma^2 = \langle x^2\rangle - \langle x\rangle^2 = \langle x^2\rangle$, and with $x = v_{\mathrm{i}}(t)$ and $y = v_{\mathrm{j}}(t+\tau)$, the cross-correlation function $R_{ij}(t)$ and the cross-correlation coefficient are related by

$$R_{ij}(\tau) = \rho\sigma^2 \tag{4.23}$$

- Bivariate Gaussian Probability Distribution:
Assume two Gaussian random variables x and y, both of zero mean, and variance σ^2. The probability $p(x, y)\, dx\, dy$ that the value of x is between x_0 and $x_0 + dx$, and that simultaneously the value of y between y_0 and $y_0 + dy$, is given by the jointly gaussian probability distribution

$$p(x_0, y_0) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left(\frac{-(x_0^2 + y_0^2 - 2\rho x_0 y_0)}{2\sigma^2(1-\rho^2)}\right) \tag{4.24}$$

In our case, the variable x is assigned to the output voltage of antenna i at time t, and y the output voltage of antenna j at time $t + \tau$.

4.5.2 Clipping correction for 4-level quantization

The following determination of the clipping correction is due to [Hagen et al 1973]:
Given two jointly normal random variables x and y with covariance μ, and given some arbitrary function $g(x, y)$, Price's theorem states that

$$\frac{\partial^m \langle g(x,y)\rangle}{\partial \mu^m} = \langle \frac{\partial^{2m} g(x,y)}{\partial x^m \partial y^m}\rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\partial^{2m} g(x,y)}{\partial x^m \partial y^m} p(x,y) dx dy \tag{4.25}$$

For random signals of zero mean, the covariance μ is identical with the cross-correlation function $R_{ij}(\tau)$ defined in Eq. 4.1. As shown by Eq. 4.1, we need to accumulate products of the voltage outputs of two antennas (i, j), but using the quantized signals rather than the continuous ones. Thus, with the identification $x = v_i(t)$ and $y = v_j(t + \tau)$, and using $\tilde{x}$ and $\tilde{y}$ for the quantized signals, we can apply Price's theorem to the 4-level cross-correlation function $R_4 = \langle \tilde{x}\tilde{y}\rangle$ such that

$$\frac{dR_4}{d\rho} = \sigma^2 \frac{dR_4}{dR} = \sigma^2 \frac{d\langle \tilde{x}\tilde{y}\rangle}{d\mu} = \sigma^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\partial \tilde{x}}{\partial x} \cdot \frac{\partial \tilde{y}}{\partial y} p(x,y) dx dy \tag{4.26}$$

($R = \rho\sigma^2$ denotes the continuous cross correlation function, for the sake of simplicity, antenna indices are omitted). The partial derivatives in the integrand are easily found by using the transfer function shown in Fig. 4.5:

$$\tilde{x} = \Theta(x) + (n-1)\left[\Theta(x - v_0) - \Theta(-x - v_0)\right]\Theta(-x) \tag{4.27}$$

where $\Theta(x) = 1$ for $x > 0$, and 0 else. Thus,

$$\frac{\partial \tilde{x}}{\partial x} = 2\delta(x) + (n-1)\left[\delta(x - v_0) + \delta(x + v_0)\right] \tag{4.28}$$

Re-writing Price's theorem, we find

$$\begin{aligned} \frac{dR_4}{d\rho} &= \sigma^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(2\delta(x) + (n-1)\left[\delta(x - v_0) + \delta(x + v_0)\right]\right) \\ &\quad \cdot \left(2\delta(y) + (n-1)\left[\delta(y - v_0) + \delta(y + v_0)\right]\right) p(x,y)dxdy\,. \end{aligned} \tag{4.29}$$

Inserting the jointly normal distribution $p(x,y)$, and evaluating the integral yields

$$\begin{aligned} \frac{dR_4}{d\rho} &= \tfrac{\sigma^2}{\pi}\tfrac{1}{\sqrt{1-r^2}}\Bigg\{(n-1)^2\left[\exp\left(\tfrac{-v_0^2}{\sigma^2(1+\rho)}\right) + \exp\left(\tfrac{-v_0^2}{\sigma^2(1-\rho)}\right)\right] \\ &\qquad +4(n-1)\exp\left(\tfrac{-v_0^2}{2\sigma^2(1-\rho^2)}\right) + 2\Bigg\}\,, \end{aligned} \tag{4.30}$$

or, alternatively, the integral form given in Eq. 4.11.

Chapter 5

LO System and Signal Transport

Robert Lucas

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

5.1 An Heterodyne Interferometer

5.1.1 The simple interferometer

This is composed of 2 antennas, a multiplier, an integrator (Fig. 5.1); we directly multiply the signals, and average in time. $\tau_G = 2\pi\mathbf{b}.\mathbf{s}/c$ is the geometrical delay. Provided the geometrical delay is compensated in the hardware, after filtering out the high frequency terms, the output of the correlator is the real part of the visibility:

$$r(t) = \mathcal{A}\cos\varphi(t) \tag{5.1}$$

A complex correlator using a quadrature network can be used to mesure the imaginary part; or (equivalently) one uses a spectral correlator.

5.1.2 The heterodyne interferometer

We now consider a more realistic two antenna system (Fig. 5.2), which includes two frequency conversions: e.g. one in the SIS mixer, and one to move the IF band to baseband for numerical sampling and digital correlation. This again is a simplification, but includes all the important effects. The PdB system has in fact 4 frequency conversions (see below).

Let us first consider the effect on phase of a simple frequency conversion.

5.1.3 Frequency conversion

The input signal to the mixer is $V(t) = E\cos(\omega t + \phi)$, and the first LO signal is $V_1(t) = E_1\cos(\omega_1 t + \varphi_1)$. Mixer output is proportional to $[V(t) + V_1(t)]^2$ and we select by a filter a band $\Delta\omega$ centered on ω_{IF}. We note: $\omega_{\mathrm{U}} = \omega_1 + \omega_{\mathrm{IF}}$, and $\omega_{\mathrm{L}} = \omega_1 - \omega_{\mathrm{IF}}$ the angular frequencies in the upper side band and in the lower

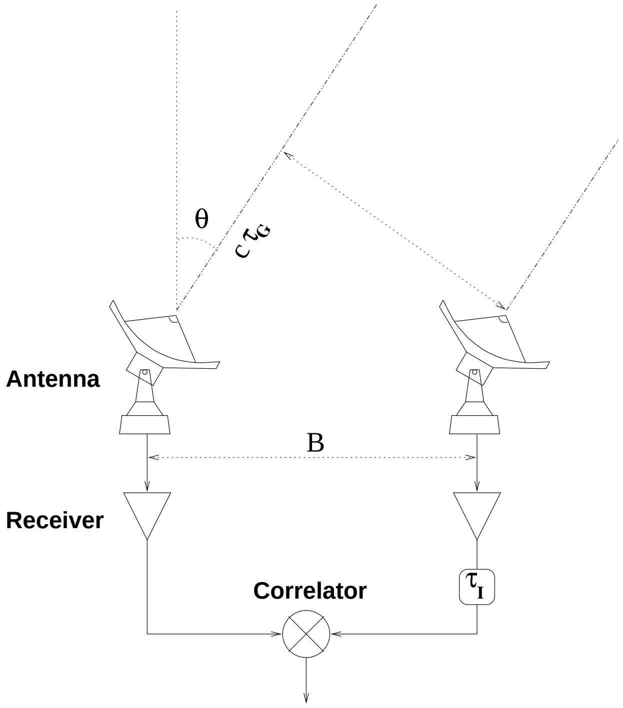


Figure 5.1: A simple, two-antenna interferometer

side band, respectively.
The IF output is

$$
\begin{aligned}
V_{\mathrm{IF}}(t) &\propto E_{\mathrm{U}} \cos\left[(\omega_{\mathrm{U}} - \omega_1)t + \varphi_{\mathrm{U}} - \varphi_1\right] + E_{\mathrm{L}} \cos\left[(-\omega_{\mathrm{L}} + \omega_1)t - \varphi_{\mathrm{L}} + \varphi_1\right] \\
V_{\mathrm{IF}}(t) &\propto E_{\mathrm{U}} \cos\left(\omega_{\mathrm{IF}} t + \varphi_{\mathrm{U}} - \varphi_1\right) + E_{\mathrm{L}} \cos\left(\omega_{\mathrm{IF}} t - \varphi_{\mathrm{L}} + \varphi_1\right) \qquad (5.2)
\end{aligned}
$$

After the frequency conversion the phase is the difference of the signal phase ant the LO phase, with a sign reversal if the conversion is lower side band:

	USB	LSB
frequency:	$\omega_{\mathrm{IF}} = \omega_{\mathrm{U}} - \omega_1$	$\omega_{\mathrm{IF}} = -\omega_{\mathrm{L}} + \omega_1$
phase:	$\varphi_{\mathrm{IF}} = \varphi_{\mathrm{U}} - \varphi_1$	$\varphi_{\mathrm{IF}} = -\varphi_{\mathrm{L}} + \varphi_1$

5.1.4 Signal phase

One antenna is affected by the geometrical delay τ_{G}, and by the phase (φ_{U} in the upper side band, φ_{L} in the lower side band), which is the quantity to be measured. We apply a compensating delay τ_{I} in the second IF, as well as a phase φ_1 to the first LO and a phase φ_2 on the second LO. We note $\Delta\tau = \tau_{\mathrm{I}} + \tau_{\mathrm{G}}$ the delay tracking error. In a 2-antenna system, we may assume that the signal path through the first antenna suffers no delay of phase offset terms. Obviously the compensating delay τ_{I} in the second antenna may need to be negative, if the second antenna is closer to the source: in that case one will apply the positive delay $-\tau_{\mathrm{I}}$ on the first antenna. In a N antenna system, one will apply phase and delay commands to all the antennas; a common delay will be applied to all the antennas since no negative delay can be built with current technology.

Let us first consider the upper side band of the first LO (second LO conversion is assumed upper side band for simplicity):

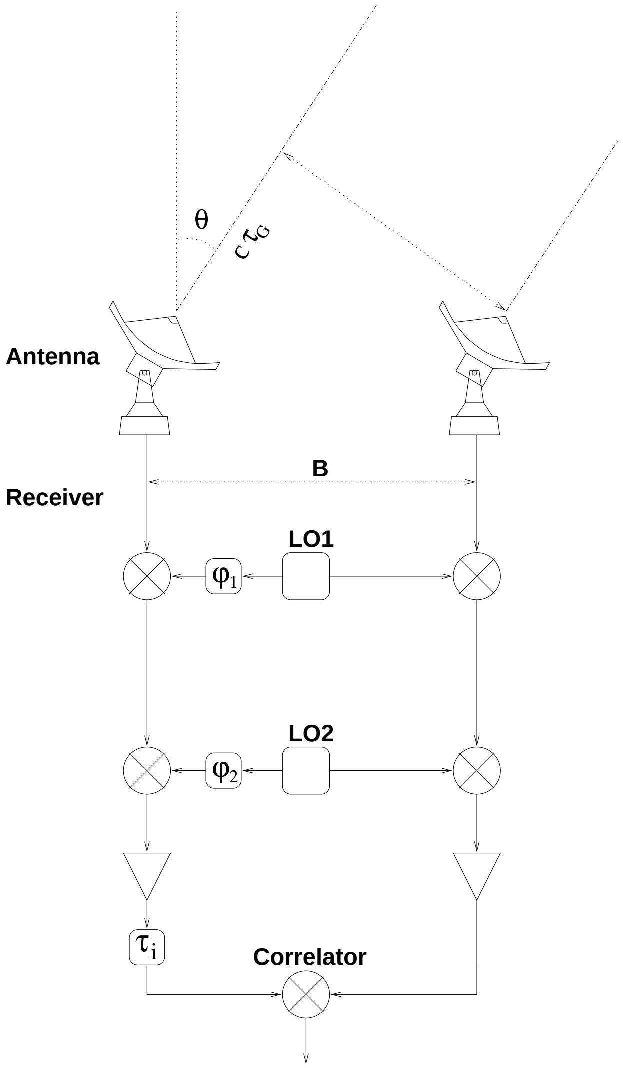


Figure 5.2: A heterodyne, two-antenna interferometer, with two frequency conversions

	USB	LSB
HF Frequency HF Phase	ω_{U} $\varphi_{\mathrm{U}} + \omega_{\mathrm{U}}\tau_{\mathrm{G}}$	ω_{L} $\varphi_{\mathrm{L}} + \omega_{\mathrm{L}}\tau_{\mathrm{G}}$
LO1 Frequency LO1 Phase	ω_1 φ_1	ω_1 φ_1
IF1 Frequency IF1 Phase	$\omega_{\mathrm{IF1}} = \omega_{\mathrm{U}} - \omega_1$ $\varphi_{\mathrm{U}} + \omega_{\mathrm{U}}\tau_{\mathrm{G}} - \varphi_1$	$\omega_{\mathrm{IF1}} = \omega_1 - \omega_{\mathrm{L}}$ $-\varphi_{\mathrm{L}} - \omega_{\mathrm{L}}\tau_{\mathrm{G}} + \varphi_1$
LO2 Frequency LO2 Phase	ω_2 φ_2	ω_2 φ_2
IF2 Frequency IF2 Phase after τ_{I}	$\omega_{\mathrm{IF2}} = \omega_{\mathrm{U}} - \omega_1 - \omega_2$ $\varphi_{\mathrm{U}} + \omega_{\mathrm{U}}\tau_{\mathrm{G}} - \varphi_1 - \varphi_2$ $\varphi_{\mathrm{U}} + \omega_{\mathrm{U}}\tau_{\mathrm{G}} - \varphi_1 - \varphi_2 + \omega_{\mathrm{IF2}}\tau_{\mathrm{I}}$	$\omega_{\mathrm{IF2}} = \omega_1 - \omega_{\mathrm{L}} - \omega_2$ $-\varphi_{\mathrm{L}} - \omega_{\mathrm{L}}\tau_{\mathrm{G}} + \varphi_1 - \varphi_2$ $-\varphi_{\mathrm{L}} - \omega_{\mathrm{L}}\tau_{\mathrm{G}} + \varphi_1 - \varphi_2 + \omega_{\mathrm{IF2}}\tau_{\mathrm{I}}$
Final	$\varphi_{\mathrm{U}} + \omega_{\mathrm{IF2}}\Delta\tau$ $-(\varphi_1 + \omega_1\tau_{\mathrm{G}})$ $-(\varphi_2 + \omega_2\tau_{\mathrm{G}})$	$-\varphi_{\mathrm{L}} + \omega_{\mathrm{IF2}}\Delta\tau$ $+(\varphi_1 + \omega_1\tau_{\mathrm{G}})$ $-(\varphi_2 + \omega_2\tau_{\mathrm{G}})$

To stop the fringes in both sidebands we need the following conditions:

$$\Delta\tau = \tau_{\mathrm{I}} + \tau_{\mathrm{G}} = 0 \tag{5.3}$$

$$\varphi_1 + \omega_1\tau_{\mathrm{G}} = 0 \tag{5.4}$$

$$\varphi_2 + \omega_2\tau_{\mathrm{G}} = 0 \tag{5.5}$$

One sees that *delay tracking* in the second IF imposes a *phase tracking* on the first and second oscillators. The delay error $\Delta\tau$ appears as a phase term proportional to frequency in the IF2 band ω_{IF2}.

The condition that e.g. $\varphi_1 = -\omega_1\tau_{\mathrm{G}}$ means that φ_1 must be commanded to vary at a rate

$$\dot{\varphi_1} = -\omega_1\dot{\tau_{\mathrm{G}}} \sim 2\pi\frac{b}{\lambda_1}\frac{2\pi}{86400} \tag{5.6}$$

which is about 10 turns per second for λ_1 = 1mm and b = 1km. The condition is much easier for the second LO. In practice the phase is commanded typically every second, as well as its rate of change during the next second (the real curve is approximated by a piecewise linear curve). Note that a linear drift with time of the phase is strictly equivalent to a small frequency offset.

5.2 Delay lines requirements

5.2.1 Single side band processing in a finite bandwidth

Assume that the conversion loss is negligible for the lower side band. At a given IF2 frequency ω_{IF2} the directly correlated signal is:

$$V_r = A\cos(\varphi + \omega_{\mathrm{IF2}}\Delta\tau) \tag{5.7}$$

while the sine correlator would give:

$$V_i = A\sin(\varphi + \omega_{\mathrm{IF2}}\Delta\tau) \tag{5.8}$$

$$V = V_r + iV_i = Ae^{i(\varphi+\omega_{\mathrm{IF2}}\Delta\tau)} \tag{5.9}$$

Assume we use a correlator with a finite bandwidth $\Delta\nu$. The correlator output is obtained by summing on frequency in the IF passband:

$$V = \int Ae^{i(\varphi+\omega_{\mathrm{IF2}}\Delta\tau)}B(\omega_{\mathrm{IF2}})d\omega_{\mathrm{IF2}} \tag{5.10}$$

where $B(\omega_{\mathrm{IF}})$ is a complex passband function characteristic of the system: gain of the amplifiers and relative phase factors.

$$V = Ae^{i\varphi} \int e^{i\omega_{\mathrm{IF2}}\Delta\tau} B(\omega_{\mathrm{IF2}}) d\omega_{\mathrm{IF2}} \tag{5.11}$$

We have assumed that the source visibility is constant across the band; the source visibility, when the delay error varies, is multiplied by the Fourier transform of the complex passband.

The delay error must be kept much smaller than the inverse of the instantaneous bandwidth to limit the signal loss to a small level. The delays are usually tracked in steps, multiples of a minimum value. To limit the loss to 1%, the minimum delay step must be $\sim 0.25/\Delta\nu$ (0.5 ns for a 500 MHz bandwidth).

5.2.2 Double side band system

In that case the signals coming from the upper and lower side bands have similar attenuation in the RF part and similar conversion loss in the mixers. They will have similar amplitudes in the correlator output. The result for the cosine correlator is:

$$V = A_{\mathrm{U}} e^{i[\varphi_{\mathrm{U}} + \omega_{\mathrm{IF2}}\Delta\tau - (\varphi_1 + \omega_1\tau_{\mathrm{G}}) - (\varphi_2 + \omega_2\tau_{\mathrm{G}})]} + A_{\mathrm{L}} e^{i[-\varphi_{\mathrm{L}} + \omega_{\mathrm{IF2}}\Delta\tau + (\varphi_1 + \omega_1\tau_{\mathrm{G}}) - (\varphi_2 + \omega_2\tau_{\mathrm{G}})]} \tag{5.12}$$

Assuming the same visibility in both sidebands:

$$V = A\cos(\varphi - \varphi_1 - \omega_1\tau_{\mathrm{G}}) e^{i(\omega_{\mathrm{IF2}}\Delta\tau - \varphi_2 - \omega_2\tau_{\mathrm{G}})} \tag{5.13}$$

If the delays are tracked, and the LO phases rotated as above, the exponential term is 1 and only the real part of the visibility is measured. Some trick is thus needed to separate the signal from the side bands.

5.3 Side band separation

The side band separation by mixer rejection is difficult for low IF frequencies, and currently works only at 3mm. The image rejection varies with frequency. There are other methods that cancel the signal in the unwanted side band by a larger factor. They are based on the fact that the LO1 phase φ_1 appears with a different sign on the USB and LSB signals.

5.3.1 Fringe rate method

One might choose to drop the phase rotation on the second LO and let the fringes drift at their natural fringe rates. These rates are opposed in sign for the USB and LSB, and they might be separated electronically. However the natural fringe rate sometimes goes to zero (when the angular distance between source and baseline direction is minimum or maximum), and at least in these cases the method would fail.

It would be more practical to offset the LO1 and LO2 phase rates $\dot{\varphi}_1$ and $\dot{\varphi}_2$ from their nominal values by the same amount ω_{OFF}. If the offsets have the same sign, they will compensate for the USB and offset the fringe rate by $2\omega_{\mathrm{OFF}}$ in the LSB. If ω_{OFF} is large enough, the LSB signal will be cancelled. Note that offsetting $\dot{\varphi}_1$ by a fixed amount is equivalent to offsetting the LO1 frequency.

This is a simple method to reject the unwanted side band. Note that the associated noise is not rejected.

5.3.2 Phase switching method

Assume a variable phase offset ψ_1 is added to the LO1 phase command appropriate for compensating the geometrical delay variation:

$$\varphi_1 = -\omega_1\tau_{\mathrm{G}} + \psi_1 \tag{5.14}$$

ψ_1 will be subtracted to the phase of the USB signal, and added to that of the LSB signal. If ψ_1 is switched between 0 and $\pi/2$, the relative phase of the USB and LSB will be switched between 0 and π, and the signals may be separated by synchronous demodulation:

ψ_1	Signal
0	$V_1 = A_U e^{i\varphi_U} + A_L e^{-i\varphi_L}$
$\pi/2$	$V_2 = A_U e^{i(\varphi_U - \pi/2)} + A_L e^{i(-\varphi_L + \pi/2)}$

Then one may compute the visibilities in each side band:

$$\begin{aligned} A_U e^{i\varphi_U} &= (V_1 + iV_2)/2 \\ \text{and} \quad A_L e^{-i\varphi_L} &= (V_1 - iV_2)/2 \end{aligned} \tag{5.15}$$

We have assumed here that we have a complex correlator (sine + cosine), or equivalently a spectral correlator measuring positive and negative delays (see Chapter 4).

One may also switch the phase by π, in which case the sign of all the correlated voltages is reversed. This has the advantage of suppressing any offsets in the system. Actually both switching cycles are combined in a 4-phase cycle:

ψ_1	Signal
0	$V_1 = A_U e^{i\varphi_U} + A_L e^{-i\varphi_L}$
$\pi/2$	$V_2 = A_U e^{i(\varphi_U - \pi/2)} + A_L e^{i(-\varphi_L + \pi/2)}$
π	$V_3 = -V_1$
$3\pi/2$	$V_4 = -V_2$

$$\begin{aligned} A_U e^{i\varphi_U} &= (V_1 + iV_2 - V_3 - iV_4)/4 \\ \text{and} \quad A_L e^{-i\varphi_L} &= (V_1 - iV_2 - V_3 + iV_4)/4 \end{aligned} \tag{5.16}$$

In a N antenna system one needs to switch the relative phases of all antenna pairs. This could be done by applying the above square-wave switching on antenna 2, then on antenna 3 at twice the switching frequency, and so on. In practice the switching waveforms are orthogonal Walsh functions.

5.4 The PdB Signal and LO transport system

A block diagram of the Plateau de Bure interferometer system is shown in Fig. 5.3.

5.4.1 Signal path

The signal path is outlined in Fig. 5.3. It shows the signal and LO paths for one antenna and one receiver band. The high frequency part (receiver) was described in Chapter 3. The amplified first IF output (1275-1775 MHz) is down-converted to the $100 - 600$ MHz band and transported to the central building in a high-quality cable. Before down-conversion, the band shape is modified by a low-pass filter; since the LO2 is at a higher frequency than the IF2, the bandpass will be reversed in the conversion, and this by anticipation compensates for the frequency dependent attenuation in the cable (which is of course higher at the high-frequency end of the bandpass).

The $100 - 600$ MHz band arriving in the central building is directed to the correlator analog IF processor inputs (with a division by 6 since there are 6 identical correlator units) and to total power detectors which are used for the atmospheric calibration and for the radiometric phase correction.

5.4.2 LO generation

The first local oscillator is a Gunn oscillator (a tripler is used for the 1.3mm receiver). The Gunn is phase-locked by mixing part of its output with a harmonic of a reference signal (used also as the second LO): the harmonic mixing produces a 100MHz signal, the phase of which is compared to a reference signal at frequency $\epsilon_1 =$ 100MHz, coming from the central building. That reference signal is used to carry the phase commands to be applied to the first LO: a continuously varying phase to compensate for earth motion and phase switching used to separate the side-bands and suppress offsets.

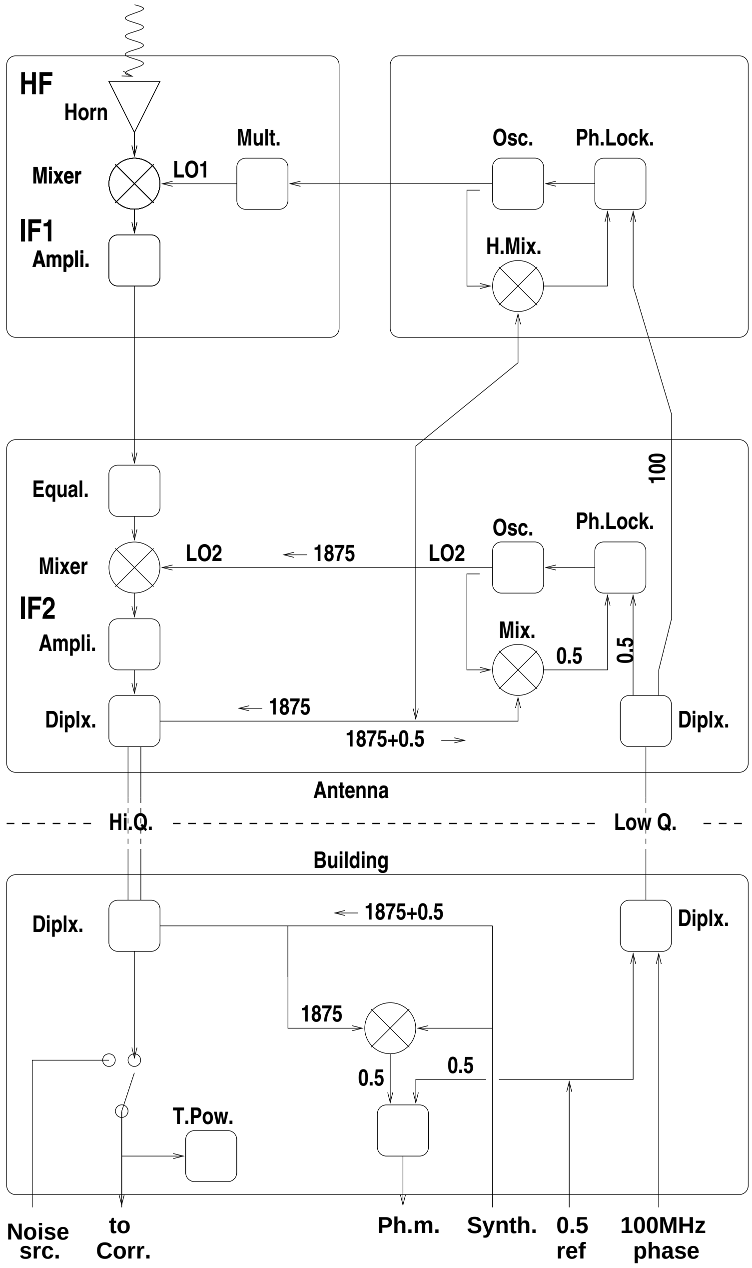


Figure 5.3: The Plateau de Bure interferometer system

The LO1 signal at ν_{LO1} may be locked either 100MHz above ("High Lock") or below ("Low Lock") the $N_{\mathrm{H}}^{\mathrm{th}}$ harmonic of the LO2 frequency ν_{LO2}:

$$\nu_{\mathrm{LO1}} = (N_{\mathrm{H}}\nu_{\mathrm{LO2}} \pm \epsilon_1)N_{\mathrm{M}} \tag{5.17}$$

The multiplication factor N_{M} is 1 for the 3mm receiver and 3 for the 1.3mm receiver.

The second local oscillator, at $\nu_{\mathrm{LO2}} = 1875 \pm 25$ MHz, is phase locked ϵ_2=0.5 MHz below the frequency sent by the synthesizer in the central building (which is under computer control and common to all antennas):

$$\nu_{\mathrm{LO2}} = \nu_{\mathrm{SYN}} - \epsilon_2 \tag{5.18}$$

The ϵ_2 reference frequency is sent to all antennas from the central building in a low quality cable, together with the ϵ_1= 100MHz reference frequency for the first LO. The ν_{SYN} is sent to the antennas via the same high-Q cable that transports the IF2 signal. No phase rotation is applied on the second local oscillator. The relation between the RF signal frequencies (in the local rest frame) in the upper and lower side bands and the signal frequency in the second IF band is thus (for high lock):

$$\nu_{\mathrm{USB}} = \nu_{\mathrm{LO1}} + (\nu_{\mathrm{LO2}} - \nu_{\mathrm{IF2}}) = (N_{\mathrm{M}}N_{\mathrm{H}} + 1)\nu_{\mathrm{LO2}} + N_{\mathrm{M}}\epsilon_1 - \nu_{\mathrm{IF2}} \tag{5.19}$$

and in the lower side band:

$$\nu_{\mathrm{LSB}} = \nu_{\mathrm{LO1}} - (\nu_{\mathrm{LO2}} - \nu_{\mathrm{IF2}}) = (N_{\mathrm{M}}N_{\mathrm{H}} - 1)\nu_{\mathrm{LO2}} + N_{\mathrm{M}}\epsilon_1 + \nu_{\mathrm{IF2}} \tag{5.20}$$

5.4.3 Further signal processing

In each correlator a variable section of the IF2 band is down-converted to baseband by means of two frequency changes, with a fixed third LO and a tunable fourth LO. It is on that LO4 that the phase rotations needed to compensate for residual phase drifts due to the geometrical delay change are applied (in fact that LO4 plays the role of the second frequency conversion in the above analysis). No phase rotation is applied on the second and third local oscillators.

The phase rotation applied on the fourth LO's is:

$$\varphi_4 = (\omega_2 + \omega_3 - \omega_4)\tau_{\mathrm{G}} \tag{5.21}$$

since the second and third conversions are LSB while the fourth is USB. It is different in the different correlator units since the ω_4 frequencies are different.

5.4.4 Phase stability requirements

Short term phase errors in the local oscillators (jitter) will cause a decorrelation of the signal and reduce the visibility amplitude by a factor

$$\eta_{12} = e^{-({\sigma_1}^2+{\sigma_2}^2)/2} = \sqrt{e^{-{\sigma_1}^2}e^{-{\sigma_2}^2}} = \sqrt{\eta_1\eta_2} \tag{5.22}$$

where σ_1 is the rms phase fluctuation of the LO in one of the antennas (σ_2 in the other). $\eta_1 = e^{-{\sigma_1}^2}$ is the decorrelation factor for one antenna; typical requirements on σ_1 are:

η_1	0.99	0.98	0.95	0.90
σ_1 (degrees)	5.75	8.1	13.0	18.5

The phase stability required on the LO2 is $\sigma_1/(N_{\mathrm{M}}N_{\mathrm{H}}) \sim 0.1°$ for a 0.95 efficiency at 1.3mm: very stable oscillators are needed.

5.4.5 Cable electrical length control

The ϵ_2 reference frequency is also used for a continuous control of the electrical length of the High-Q cables transporting the IF2 signal from the antennas to the correlator room in the central building. A variation ΔL in the electrical length of the High-Q cable will affect the signal phase by $360\Delta L/\lambda_{\rm IF2}$; for a length of 500m and a temperature coefficient of 10^{-5} we have a variation in length of 5mm or 17ps, which translates into a phase shift of 4 degrees at the high end of the passband: this is a very small effect.

The same length variation induces a phase shift of $360 \times 0.017 \times 1.875 = 11.5$ degrees at the LO2 frequency. This signal being multiplied by $(N_{\rm H}+1)N_{\rm M} \sim \nu_{\rm USB}/\nu_{\rm LO2} \sim 120$ for the 1.3mm receiver, we have a totally unacceptable shift of about 4 turns. The cables are buried in the ground for most of their length; however they also run up the antennas and suffer from varying torsions when the sources are tracked, and in particular when the antenna is moved from the source to a phase calibrator.

For this reason the electrical length of the cables is under permanent control. The LO2 signal is sent back to the central building in the High Q cable, and there it is mixed with the $\nu_{\rm LO2} + \epsilon_2$ signal from the synthesizer. The phasemeter measures every second the phase difference between the beat signal at 0.5 MHz and a reference 0.5 MHz signal.

The measured phase difference is twice the phase offset affecting the LO2, it is used by the computer to correct the LO1 phase φ_1 after multiplication by $\nu_{\rm LO1}/\nu_{\rm LO2}$.

5.5 Next generation instruments

Next generation instruments will operate at higher frequencies, and need higher bandwidths, and better angular resolution. The major changes expected are:

- Use of optical fibers rather than cables. Actually this is already the case in some interferometers.
- Digitize higher in the signal chain. transporting digital signals require more bandwidth but is more accurate.
- Possibly generate LO signals using infrared lasers rather than by multiplying lower frequency signals.

Chapter 6

The Plateau de Bure Interferometer

Roberto Neri

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

6.1 History

The design of the millimeter wave interferometer started in June 1979, in the year of the foundation of IRAM, the Institut de Radioastronomie Millimétrique. The construction of the first antenna was completed in June 1987, and three years later an interferometer consisting of three antennas was opened to guest observers. First fringes at 230 GHz were obtained in April 1995, and the current configuration of the interferometer was attained one year later. Work is currently in progress to extend the north-south track and for the construction of the sixth antenna. Starting with the commissioning period in 1990 up to the end of 1998, the Plateau de Bure interferometer was able to carry out more than 450 different projects which involved more than 200 investigators from all around the world.

6.2 Description

The Plateau de Bure interferometer is located in the South of the French Alpes, near St-Etienne en Dévoluy in the Départment of Haute Alpes. The interferometer's altitude is 2552 m (2560 m at the intersection of the azimuth and elevation axes of the telescopes) and its longitude and latitude at the array phase center are 05:54:28.5 E and 44:38:02.0 N. A cable car provides access to Plateau de Bure. A hangar in which the sixth antenna is currently under construction, is used for antenna maintenance, overhaul periods and, in general, for antenna repair work which needs to be carried out in between times but under safety conditions. The hangar houses a few workshops for cryogenics, mechanics, electronics, a power station which provides electric autonomy in case of interruptions in the external power supply and finally, the correlator room and the control room for remote array operation. Almost adjacent to the hall are the living quarters for the staff who supports the uninterrupted round the clock operation of the interferometer.

Currently, the interferometer consists of five antennas arranged in a T–shaped pattern extending over a maximum of 408 meters east-west and 232 meters north-south. A sixth antenna is expected to be ready mid 2000, and further baseline extension of the north track is under way.

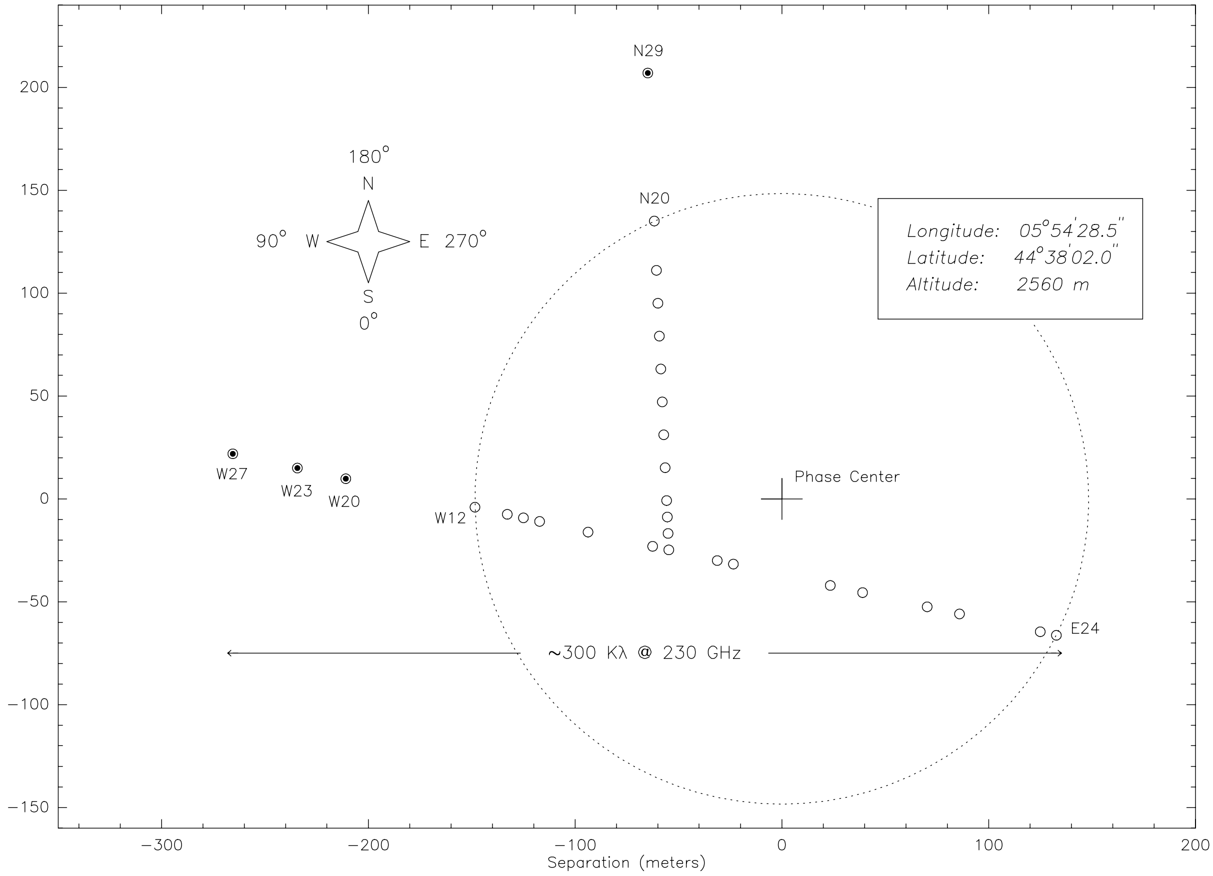


Figure 6.1: PdB interferometer station layout as of September 1998. The interferometer origin is defined as the center of the circle which goes through the stations N20, E24 and W12, the so-called IRAM phase center, and thus gives a unique vector definition to each station.

The antennas are conceptually identical: they all have a fully steerable alt-az mount which incorporates a self-propelled transporter for moving the antennas (130 tons) along the tracks between stations. Each antenna is a 15 m diameter Cassegrain telescope with the backstructure and quadrupod legs largely made of carbon fiber for high thermal stability. The high precision of the reflecting antenna surface (40-60μm) guarantees best performance: all antennas have essentially the same sensitivity (22 Jy.K^{-1} at 3 mm, 35 Jy.K^{-1} at 1 mm – see lecture by A.Dutrey), and very similar pointing and focusing characteristics.

All the antennas are equipped with dual-frequency SIS receivers operating simultaneously in the 82 GHz−115 GHz and 205 GHz−245 GHz range. Typical double sideband receiver noise temperatures are between 25 K and 50 K at 3 mm and between 40 K and 60 K at 1 mm. The receivers upper and lower sidebands are separated by the correlators with a rejection better than 26 dB. The lower to upper sideband gain ratio depends on the receiver and varies typically between 0.2 and 4.0 under standard operating conditions in the 3 mm band, and yields essentially a double-side band tuning in the 1 mm band. Pure single sideband tuning (with rejection 15 to 25 dB) is also possible in the 3 mm band, with receiver temperatures around 60 to 80 K.

Six totally independent correlators units are available that provide an 87% correlation efficiency (for more details see H.Wiesemeyer's lecture). Each correlator unit provides 4 choices of bandwidth/channel configurations, a single unit provides an additional configuration with a nominal velocity resolution of 50 m.s^{-1} at 230 GHz. The correlators can independently be connected either to the 3 mm (100-600 MHz IF1) or to the 1 mm receivers (100-600 MHz IF2).

A central control computer coordinates the entire interferometer (antennas, receivers and correlators and quite some other equipment) and makes the data acquisition. Raw data corresponding to the individual dumps of the correlator buffers will not be available as real-time jobs apply automatic calibrations (clipping correction, apodization, FFT, sideband separation, small delay correction, bandpass correction and other corrections) and make automatic data quality assessments (marking bad data, shadowing, phase lock, just to cite a few flags) before data is written to disk. A second workstation provides the software resources for offline data analysis and for data archiving before transfer to the Grenoble headquarters.

6.3 Array operation

6.3.1 Array calibration

The astronomical setup of the interferometer involves a number of steps that are done under the joint responsibility of the array operator and of the astronomer on duty (AoD). The goal of the setup is to maximize the interferometer performance in view of sensitivity and positional precision.

Change of array configuration

A change of configuration is the responsibility of the operators and of the technical staff. Since most projects, as mapping, mosaicing and snapshot observations, require more *uv*–coverage than a single configuration can provide, the antennas are moved typically every three weeks or so, to a new configuration. Every additional configuration increases the mapping sensitivity and the uniformity of the *uv*–coverage by adding $N(N-1)/2$ baselines to the sampling function (these are 10 baselines during the winter period, 6 baselines during the summer period when the array is operated with only 4 antennas). Configurations are usually selected among six types according to several criteria: antenna availability, project type, atmospheric seeing, *uv*–coverage, pressure in local sidereal time, sun avoidance and other factors.

Six primary configurations are needed to cover the desired range of angular resolution at the two operating frequencies with 5 antennas:

Configuration	Stations
D	W05 W00 E03 N05 N09
C1	W05 W01 E10 N07 N13
C2	W12 W09 E10 N05 N15
B1	W12 E18 E23 N13 N20
B2	W23 W12 E12 N17 N29
A	W27 W23 E16 E24 N29

The configurations can be combined to produce five sets of configurations for different angular resolution:

Set	Configurations	Purpose
D	D	detection / lowest resolution
CD	D, C2 or C1	$3.5''$ at 100 GHz
CC	C1, C2	higher resolution than CD
BC	B1, C2	$2.0''$ at 100 GHz
BB	B1, B2, C2	higher resolution and sensitivity
AB	A, B1, B2	$1.0''$ at 100 GHz

Special configurations and sets of configurations are used during the annual antenna maintenance period which is usually between May and October. During this period observations at 1 mm are for most of the time not feasible, specially in the two extended B configurations. Observations in the A configuration whether at 3 mm or 1 mm will in general only be scheduled during the winter period. Requested non-standard configurations are considered only in exceptional cases.

Antenna focus

Sensitivity is one of the most important concerns. As a rule of thumb, an axial displacement of the secondary by $\sim \lambda/3$ results in a 20% loss of sensitivity. To avoid losses larger than 3%, the position of the secondary needs to be measured to much better than $\lambda/10$ on regular time intervals. The positional precision, however, depends on the source strength, the operating wavelength, the sampling of secondary positions and, finally, on atmosphere stability. In general, the focus is measured at 3 mm on a strong quasar by displacing the secondary in steps of 1 mm (in steps of 0.45 mm if done at 1 mm). This is systematically done by the operators at the beginning of every project and is automatically verified by the system every hour during project execution.

Antenna pointing

A high pointing accuracy is demanded in view of sensitivity and mapping quality. Antenna pointing errors affect the global sensitivity of the interferometer and may lead to severe errors in the image restoration process. As a rule, a pointing precision of $\Delta\theta \sim \theta_{\mathrm{FHWP}}/20$ is desirable at the highest frequency. The good pointing accuracy results from an optimized structural design: a good knowledge of the gravitational load, a good positional stability of the receivers (a good alignment is needed for dual-frequency observations), a precise control of the secondary, high precision bearings and position encoders, a good servo system, ... and a good software control for repeatable antenna pointing errors. The quality of a pointing model is generally limited by wind and thermal load effects. The absolute pointing accuracy achievable with the IRAM antennas is in general below the 2-3$''$ rms at each axis with a slightly higher uncertainty in elevation. Such a pointing accuracy leads to very small intensity variations, most of the time with negligible effects on the image reconstruction. Higher accuracy is obtained by regular relative pointing measurements every hour.

Each antenna is characterized by a fixed set of pointing parameters. These are measured only in certain circumstances: when an antenna is going to see first light, when modifications are made which may affect the pointing of an antenna, or more generally in cases of suspected pointing problems. In these cases a precise interferometric pointing session, eventually with a preceding less sensitive full-sky single-dish session, is required to derive the full set of antenna pointing parameters. Such pointing sessions are reduced with a dedicated non-linear fitting program in use at Plateau de Bure.

The pointing model is actually based on 5 parameters only, all others being negligibly small. These parameters are: IAZ and IEL (the azimuth and elevation encoder zero point correction), COH (the antenna horizontal collimation), and IVE and IVN (the antenna East-West and North-South inclination). IAZ, IEL, IVE and IVN are in station dependent, while COH is in principle an antenna constant. IAZ, IEL and COH are measured in interferometric mode by pointing on a few low elevation and high elevation sources. In general, three strong quasars at 3 mm are fully sufficient. The remaining two parameters, IVE and IVN are measured on every project start with an inclinometer by making an antenna turn through 360°.

Delay measurements

Delay measurements aim at the correction of cable length (electric path) differences between two antennas after compensation of the geometrical path length. An improper knowledge of the difference in cable length is visible as a frequency dependent phase slope in the intermediate frequency bands (IF1 and IF2), and, depending on the amplitude of the slope, may result in a more or less important loss of sensitivity. The delay is measured by a cross-correlation on a strong radio source at the beginning of every project.

Baseline lengths measurements

The goal is to measure the position of each antenna i relative to a common reference point (distances X_{ij}, Y_{ij}, Z_{ij} between antennas i and j or distances dX_i, dY_i, dZ_i with respect to the theoretical station position) in order to subtract the phase term $2\pi w$ (see lecture 2 by S.Guilloteau) at any hour angle and declination from the observed phase. The absence of a good baseline solution is equivalent to having large uncertainties in the baseline separation between different antennas. As a consequence, the geometrical delay might improperly be compensated and large time-variable phase errors might affect observations.

Though the quality of a baseline solution is easily found out – the calibrator's visibility phase shouldn't vary with reference to the phase tracking center as function of hour angle and declination – a good baseline solution is truly indispensable for the purpose of phase calibration. Phase errors can often be more deleterious on compact configurations where source visibilities are stronger than on extended configurations. As a reference, winter conditions allow baselines in the D configuration to be measured at 3 mm with a $5° - 8°$ phase accuracy and with $5° - 20°$ in the A configuration. In summer conditions the accuracy is often $2 - 3$ times lower.

Though no high accuracy is needed for antenna positioning (offset position from the target location is routinely within a wavelength), the actual antenna position has to be known with high precision: within a small fraction of a wavelength (70-300μm). The precision is limited essentially by the atmosphere and by thermal effects.

The baseline parameters can be obtained to high accuracy from observations of a number k of relatively strong point sources, well-distributed in hour-angle and declination, for which accurate positions are available. The analysis of these observations is usually carried out with CLIC, the calibration program, using a least-square-fit analysis on the geometric phase difference for antenna pairs (i, j):

$$\phi^g_{ij} = \phi^s_{ij} + \phi^a_{ij} = 2\pi\, w =$$

$$= \frac{2\pi}{\lambda}\, \underbrace{(X_{ij}, Y_{ij}, Z_{ij})}_{b} \cdot \underbrace{\begin{pmatrix} cosH cos\delta \\ -sinH cos\delta \\ sin\delta \end{pmatrix}}_{s} + \phi^a_{ij}$$

$$\longrightarrow \quad \Delta\phi^g_{ijk} = \frac{2\pi}{\lambda}\,(\Delta b_{ij} \cdot s_k + \underbrace{b_{ij} \cdot \Delta s_k}_{\simeq 0}) + \Delta\phi^a_{ijk}$$

where ϕ^s_{ij} is the assumed geometrical phase between the two stations, H and δ the hour angle and declination of the source, and where $\Delta\phi^a_{ijk} = O_{ij} \cos \mathrm{El}_k$ are elevation dependent correction terms for the non-intersection of the elevation and azimuth axes in the nodal point of the antennas. These terms are well-known and stand for non-negligible elevation dependent variations of the visibility phase which need to be removed as accurately as possible before solving for the baselines.

In theory, three sources are sufficient to measure the actual baseline lengths, in practice 10-12 sources are necessary to obtain an accurate measurement. Since a displacement by $1''$ at 100 GHz on a baseline of 100 m translates already to a phase offset of $\sim 58.2°$ (~ 1 rad), the positions of the radio sources used for baseline measurements need to be known with an accuracy Δs_k better than $0.02''$.

The baseline equation implies that positional errors are equivalent to phase errors. Since baseline length errors scale with the angular separation between calibrator and source, the aim is to have calibrators as close as possible to minimize the phase errors.

Sometimes, accurate baselines are not required as in the case of self-calibration projects. Sometimes, however, even if good baselines are required, they simply cannot be determined precisely enough after a change of configuration. Projects observed in the meantime will then need to wait for a better baseline model. Such projects will in general not be phase-calibrated by the astronomer on duty, but phase-calibration has to be done later on by the proposers of the observations.

Gain measurements

Gain measurements (GAIN scans) are cross-correlations on strong radio sources which are essentially used to measure the image to signal sideband ratios for both the 3 mm and 1.3 mm receivers. The required sideband ratio depends on the project, the achievable sideband ratio depends on the receiver and the frequency. An accurate measurement of the receiver gain is necessary for a good estimate of the atmospheric opacity and of the associated thermal noise with which the atmosphere contributes during the observations. Therefore, results of a gain measurement are followed by an atmospheric calibration (scan `CALI`).

Receiver stability

As a rule, a high receiver stability $\leq 3.10^{-4}$ is never required. Sometimes, however, depending on atmospheric conditions, array configuration and observing frequency, a higher stability may be desirable in view of a very promising radiometric phase correction. Though such a high stability is not always achievable on all the receivers, it makes possible an improvement in data quality when the atmospheric phase correction technique becomes practical (see lecture 9 by M.Bremer). Experience at Bure from the last three years shows that the radiometric phase correction is quite efficient under clear sky conditions: from spring to autumn essentially during the evening and morning hours, in winter almost always when the weather allows to observe.

Since observations on more compact baselines suffer less from the effects of the atmospheric phase noise – for reference, an rms of less than 10° rms at 3 mm is routinely obtained on the shortest baselines – a high receiver stability in compact configurations is only exceptionally required. Typically, under average observing conditions with a receiver stability of 3.10^{-4} we may already correct atmospheric phase fluctuations with a precision of 10° at 115 GHz.

6.3.2 Array observations

Setting up a project

Since projects are spread over typically a few months, it is impractical that astronomers actually come to the interferometer for their observations. In some exceptional case, however, when observations require rapid decisions, the presence of a visiting astronomer may become necessary. Up to now and after almost ten years of operation, only a handful of projects required the presence of a visiting astronomer. Only non-standard observations like mapping of fast moving objects, coordinated observations may require a member of the project team to be present on the site. All observations are carried out "in absentee", and a *local contact* is assigned to each project.

The observer has to specify all aspects of his/her program in an observing procedure. For routine observations, this is usually done with the help of the local contact by parameterizing the general observational procedure. Once the procedure is written, a copy is made available to the operation center at Plateau de Bure. Before start, further verifications will be made by the scientific coordinator and, to finalize the procedure, by the astronomer on-duty who makes a last check by looking at the technical details in the proposal, at the technical report and at the recommendations made by the programme committee.

Quite some time, however, may pass between the preparation of an observational procedure and the actual observations. Depending on the requirements, between a few hours and a few months may go for the decision to start the observations. On average 90% of the projects are completed within 6 months from their acceptance.

Observations

For the observations, the array is operated by an operator with the assistance of an astronomer and under the supervision of the scientific coordinator. The operator has the full responsibility for conducting all observations following pre-established observing procedures or with the help of the astronomer in case of unpredicted events.

The operator will execute the observing procedure according to a pre-established planning which allows for some flexibility in the scheduling, and to a few criteria (as the maximum amount of precipitable water in the atmosphere, the required atmospheric phase stability, the requested observing frequencies, the declination of the targets, the sun avoidance limit and a few other aspects) which will help both the operator and the astronomer in their final decision-making on which project to carry out as next. As a rule, excellent atmospheric conditions will be used for high grade projects requesting sensitivity at high frequencies while the remaining time will in general be devoted to projects which require less stringent atmospheric conditions.

Once a project is selected, the operator will start the observing procedure which sets up the needed equipment configuration (essentially sky frequencies, correlator settings and target coordinates according with the observer's wishes) and will start preparing the interferometer for the observations: the receivers are

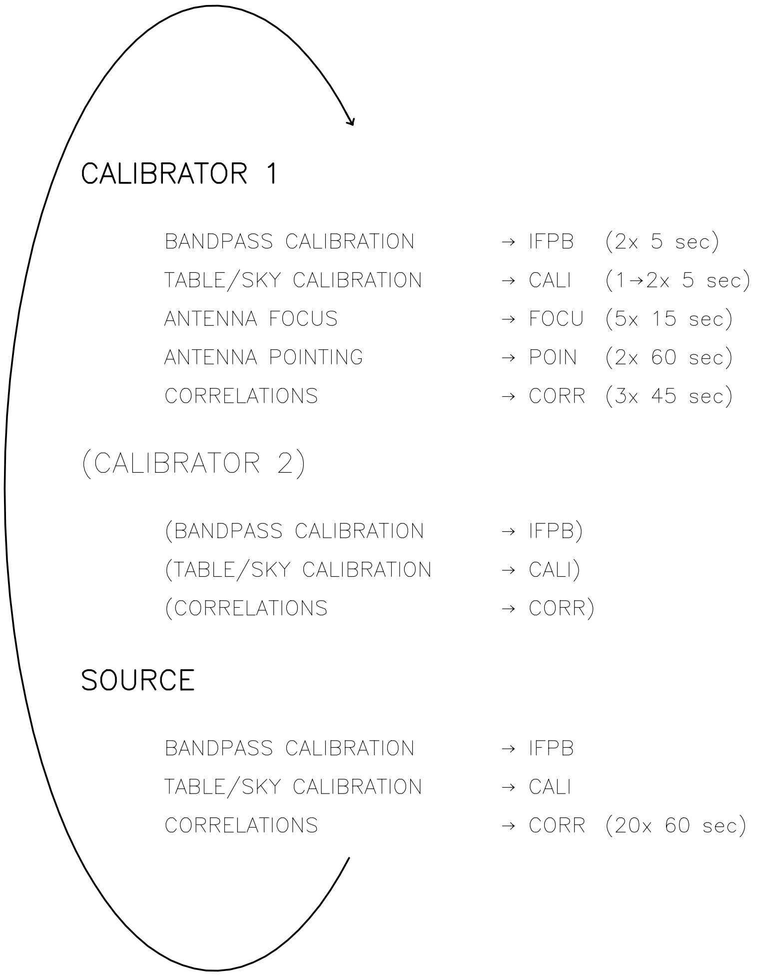


Figure 6.2: Standard observations: a cyclic sequence of measurements. `IFPB` scans aim at calibrating the IF passband, `CALI` are auto-correlations on a hot load (290 K) and on the sky (on a 15 K load only at the beginning of every project) to adjust the temperature response of the array, `FOCU` and `POIN` scans measure focusing and pointing offsets (done only every 45 minutes) and `CORR` scans are cross-correlations (complex visibilities). In projects requesting two calibrators every other calibration is made on the second calibrator.

tuned, the gains and the zero-delay of the receiving antenna is adjusted and verified, the antenna pointed and focused, the RF passband is measured and the temperature scale of the interferometer calibrated. The flux of the primary calibrators are then verified, eventually replaced if their flux density has dropped too much, and the observations started.

As soon a project is started, the astronomer on-duty will monitor the execution of the project and the data quality by examining the visibility amplitude and phase of the calibration sources, the antenna tracking in presence of wind, the antenna pointing corrections, and all time-dependent instrumental and atmospheric parameters which could have some implications on the observations. Furthermore, to avoid further observing on a target with wrong coordinates, the astronomer will verify the presence or absence of line and/or continuum emission according to the expected values quoted in the proposal. Finally, the astronomer on-duty will provide pre-calibrated data on a best effort basis. Depending on project complexity and needs, further data analysis is sometimes required on the site to decide on follow-up observations.

When the observations are running, commands are regularly issued to the antennas and to the peripheral equipment (phase rotators, correlators and others) following a well-defined, cyclic sequence as shown in Figure 6.2. This sequence may slightly change depending on the number of calibrators and on the number of phase centers (i.e. the fields of view requested for different sources or for mosaic-type observations) the observer wishes to track in a single run. Typical observations at Plateau de Bure fall in one of the following categories:

- Detection: deep integrations aiming at measuring the flux densities of faint targets (continuum emission and/or line emission/absorption) with the interferometer mostly in compact configurations.
- Snapshot: observations in one or more configurations, aiming at measuring apparent sizes in a small sample of targets, in some cases even allowing for mapping. The observational procedure sets up short integrations on individual objects in a cyclic manner.
- Mapping: observations generally in two or more configurations aiming at mapping a single object.
- Mosaic: similar to snapshot observations, except that the array switches between adjacent, half-beam spaced phase centers to map field of views which are more extended than the primary beam of the antennas.

Monitoring project execution

Under normal circumstances only a few parameters of interest are regularly verified and corrected (mostly automatically) during the observations, but instantaneous (every second) and much more detailed information can be obtained at any time by connecting to the equipment (receivers, antenna control parameters, digital correlator units and others). During the operation the array status is continuously monitored so that the operator can provide fast feedback in response, at any time when necessary. An automatic data quality assessments (flagging bad data, antenna shadowing, receiver phase lock and others) before writing data to disk. The astronomer on-duty has the responsibility of periodically monitoring the data acquisition and to write a few notes assessing the data quality during and after the observations. Monitoring the progress of a project by making intermediate data reductions, however, is the responsibility of the observer. This is not the responsibility neither of the astronomer on-duty nor of his/her local contact.

6.4 Proposal submission and contact people

Quite some people are required to run such a complex instrument as an interferometer. Sooner or later you will meet some of these people, but for most of the projects only a few will play an essential role.

At proposal submission time, you will first get in touch with the scientific secretary who will address you a confirmation of the proposal reception shortly after the deadline. Once a proposal is registered it is sent to the members of the Programme Committee (eight members: two from each country of the IRAM partner organizations plus two external) and, at the same time, its feasibility assessed by the scientific coordinator. Technical considerations as observing strategy, observational risk factors and other issues are communicated to the Programme Committee, only if necessary and only at the time of the meeting to avoid any technical remark to influence the scientific evaluation. Shortly after the meeting, during which the Programme Committee expresses its recommendations, a global proposal evaluation is made by the IRAM director who takes the final decision. Thereupon, a notification is addressed to you which contains the recommendations of the Programme Committee and a technical report. If the proposal is not rejected from the beginning, a local contact (a staff astronomer) will be appointed to the project and his name communicated to you.

In the course of the observations, only four persons will play a role for the principal investigator who proposed the observations: the local contact, the scientific coordinator, the astronomer on-duty and the array operator. The local contact, who is the direct interlocutor of the observer, is a staff astronomer whose role is to help the observer in a concerted effort to prepare his/her observations and, later on, in the Grenoble headquarters to help (if needed) the observer in calibrating the data.

Finally, once the observations are completed and before coming to Grenoble, the principal investigator or one of his team members will need to get in touch with two persons: the local contact and the coordinator for the data reduction activities who will finalize the stay of the visiting astronomer at the Grenoble headquarters.

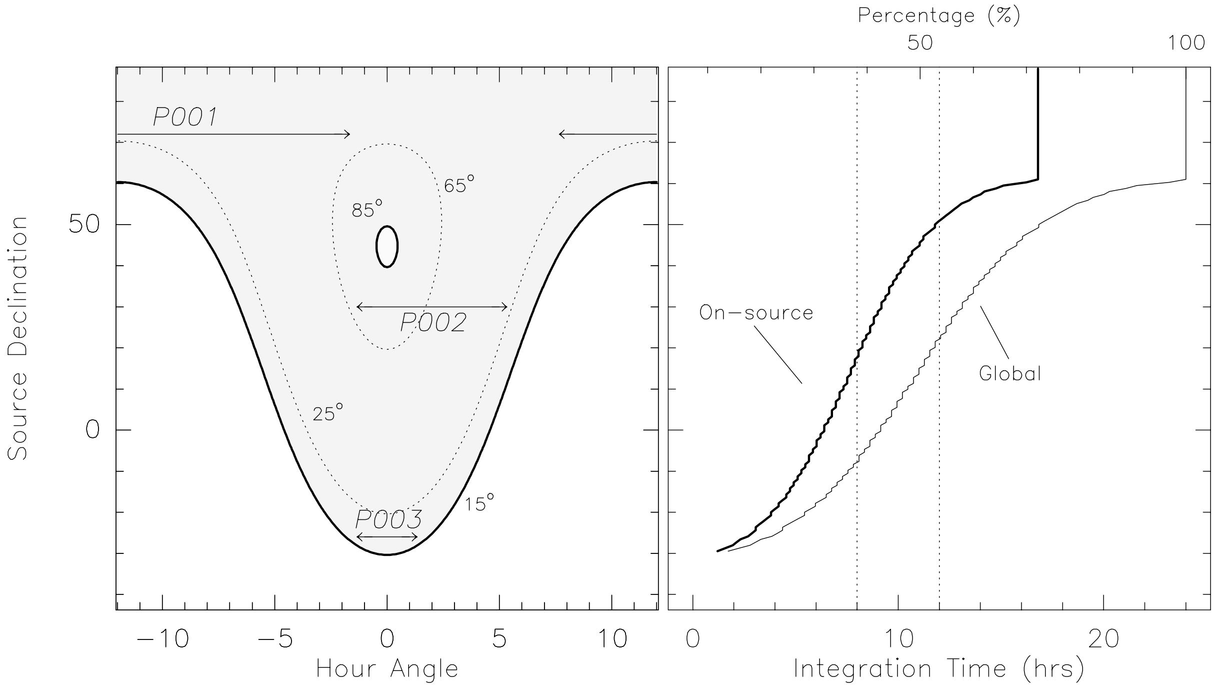


Figure 6.3: Source visibility and effective observing time. The scheduling priority is in general inversely proportional to the declination (observations of low declination sources tend to be more difficult, as they cannot be carried out at any time – the shaded region is the sky above 15° elevation). The total observing time per track for depends on the declination of the source and is usually limited to 8-12 hrs for sources of the northern hemisphere. For standard mapping projects, observational overhead counts in by 28% of the time. This is equivalent to an effective on-source integration time of about 6-9 hrs.

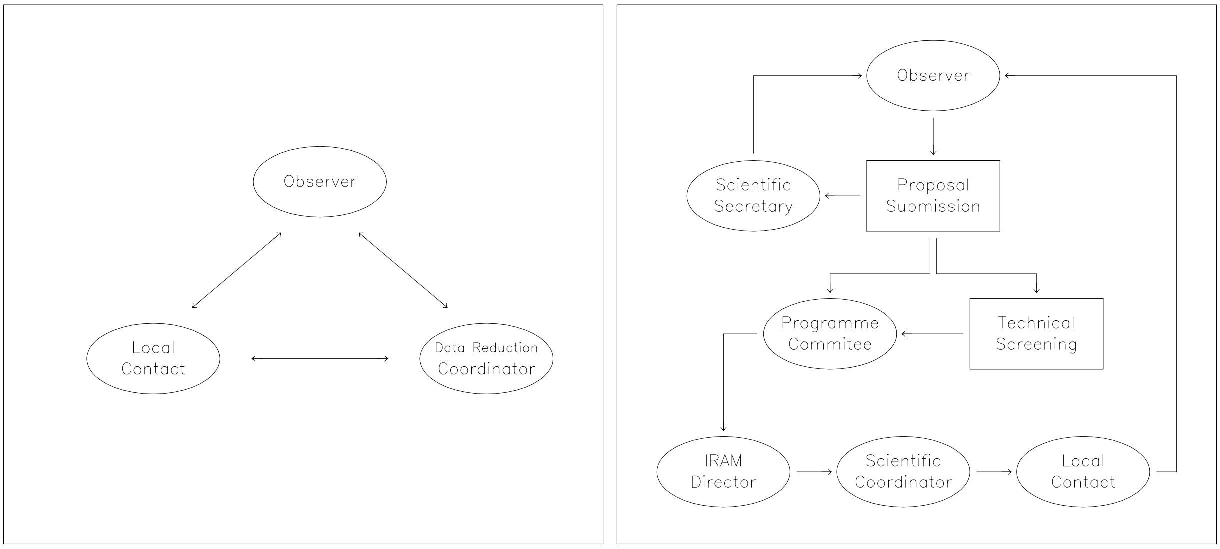


Figure 6.4: Proposal submission and contact persons at IRAM

Chapter 7

Band pass and Phase Calibration

Robert Lucas

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

7.1 Definitions and formalism

As has been seen in the previous lectures, each interferometer baseline provides a measurement of the source visibility at a given point in the u, v plane of spatial frequencies; the source brightness distribution can then be reconstructed by an appropriate Fourier Transform.

In reality things are not so simple. Interferometers are designed with a lot of care; however many electronic components will have variable gains both in amplitude and in phase; these variations will affect the results and have to be taken out. It is generally sometimes more efficient to have a slightly varying instrument response, and a more sensitive instrument, than a very stable one with less sensitivity, provided the varying terms in the response are slow and may be easily calibrated out. At millimeter wavelengths the atmospheric absorption and path length fluctuations will dominate the instrument imperfections in most cases.

For a given observation, if we interpret the correlator response (amplitude and phase) as the source visibility, ignoring any imperfections, we have an observed (apparent) visibility $\widetilde{V}_{ij}(t)$, where i, j are antenna numbers, ν the frequency and t is time. If the true source visibility is $V_{ij}(t)$, we may define :

$$\widetilde{V}_{ij}(t) = \mathcal{G}_{ij}(t)V_{ij}(t) + \epsilon_{ij}(t) + \eta_{ij}(t) \tag{7.1}$$

where the $\mathcal{G}_{ij}(\nu, t)$ are the complex gains of each baseline. $\eta_{ij}(t)$ is a noise term resulting from thermal fluctuations in the receivers; $\epsilon_{ij}(t)$ is an offset term. This assumes that the system is linear. π phase switching applied on the first local oscillators is a very efficient method of suppressing the offsets $\epsilon_{ij}(t)$; they are generally negligible and will not be considered any further.

7.1.1 Baseline based vs antenna based gains

Since amplitude and phase distortions have different physical origins it is generally useful to write

$$\mathcal{G}_{ij}(t) = g_i(t)g_j^*(t) = a_i(t)a_j(t)e^{i(\phi_i(t)-\phi_j(t))} \tag{7.2}$$

Here we have split the gains into antenna based factors. This is generally legitimate since the gains represent properties of the data acquisition chains which are in the analogue part of the system. The correlator itself is a digital machine and we assume it is perfectly working (including the clipping correction). This assumption is certainly valid when considering a single frequency and a single instant. When we start averaging in time or frequency, the average of the product may not be the product of averages, and we may have some baseline-based effects.

The baseline-based gains can be determined by observing a point source. This is usually a strong quasar. In that case the true visibilities $V_{ij}(t)$ should all be equal to the quasar flux density S. Then

$$\mathcal{G}_{ij}(t) = \frac{\widetilde{V}_{ij}(t)}{S} \tag{7.3}$$

The antenna gains $g_i(t)$ can also be deduced from the non-linear set of equations:

$$g_i(t)g_j^*(t) = \frac{\widetilde{V}_{ij}(t)}{S} \tag{7.4}$$

This is a system with N complex unknowns and $N(N-1)/2$ equations. In terms of real quantities there are $N(N-1)$ measured values (amplitudes and phases; there are only $2N-1$ unknowns since one may add a phase factor to all complex gains without affecting the baseline-based complex gains. When N is larger than 2 the system is over determined and may be solved by a method of least squares.

If we note $\widetilde{V}_{ij} = \tilde{A}_{ij}e^{i\tilde{\varphi}_{ij}}$, the equations for phases are simply:

$$\phi_i - \phi_j = \tilde{\varphi}_{ij} \tag{7.5}$$

It can be shown that the least-squares solutions (when the same weight is given to all baselines, and if we impose the condition $\sum_{j=1,N} \phi_J = 0$), is given by:

$$\phi_i = \frac{1}{N}\sum_{j\neq i} \tilde{\varphi}_{ij} \tag{7.6}$$

For the amplitudes we can define in order to get a linear system:

$$\gamma_i = \log g_i, \qquad \tilde{\alpha}_{ij} = \log \tilde{A}_{ij} \tag{7.7}$$

$$\gamma_i + \gamma_j = \tilde{\alpha}_{ij} \tag{7.8}$$

This time the least square solution is, when the same weight is given to all baselines:

$$\gamma_i = \frac{1}{N-1}\sum_{j\neq i} \alpha_{ij} - \frac{1}{(N-1)(N-2)}\sum_{j\neq i}\sum_{k\neq i,>j} \alpha_{jk} \tag{7.9}$$

Obviously this antenna gain determination needs at least three antennas. For three antennas it reduces to the obvious result:

$$g_1 = \frac{\tilde{A}_{12}\tilde{A}_{13}}{\tilde{A}_{23}} \tag{7.10}$$

These formulas can be generalized to the cases where the baselines have different weights.

It can be seen in the above formulas that the precision to which the antenna phases and amplitudes is determined is improved by a factor $\sqrt{N}$ over the precision of the measurement of the baseline amplitudes and phases.

7.1.2 Gain corrections

The determination of antenna-based gains (amplitudes and phases) has an obvious advantage: the physical cause of the gain variations are truly antenna-based. One may solve for the gains at the time of the observations, and correct the occurring problems to improve the quality of the data. One may re-point or re-focus the antennas to correct for an amplitude loss, correct for an instrumental delay (affecting the frequency dependence of the phases) ...

7.2 Bandpass calibration

In the previous section we have we have considered a monochromatic system. We have actually finite bandwidths and in principle the gain coefficients are functions of both frequency and time. We thus write:

$$\widetilde{V}_{ij}(\nu,t) = \mathcal{G}_{ij}(\nu,t)V_{ij}(\nu,t) = g_i(\nu,t)g_j^*(\nu,t)V_{ij}(\nu,t) \tag{7.11}$$

If we make the assumption that the passband shape does not change with time, then we should have for each complex baseline-based gain:

$$\mathcal{G}_{ij}(\nu,t) = \mathcal{G}_{\mathrm{B}\,ij}(\nu)\mathcal{G}_{\mathrm{C}\,ij}(t) \tag{7.12}$$

The same decomposition can also be done for the antenna-based gains:

$$g_i(\nu,t) = g_{\mathrm{B}\,i}(\nu)g_{\mathrm{C}\,i}(t) \tag{7.13}$$

$g_{\mathrm{B}\,i}(\nu)$ is the antenna complex passband shape, which by convention we normalize so that its integral over the observed bandpass is unity; then $g_{\mathrm{C}\,i}(t)$ describes the time variation of the complex gains.

Frequency dependence of the gains occurs at several stages in the acquisition chain. In the correlator itself the anti-aliasing filters have to be very steep at the edges of each sub band. A consequence is that the phase slopes can be high there too. Any non-compensated delay offset in the IF can also be seen as a phase linearly dependent on frequency. The attenuation in the cables strongly depends on IF frequency, although this is normally compensated for, to first order, in a well-designed system. The receiver itself has a frequency dependent response both in amplitude and phase, due the IF amplifiers, the frequency dependence on the mixer conversion loss. Antenna chromatism may also be important. Finally the atmosphere itself may have some chromatic behavior, if we operate in the vicinity of a strong line (e.g. O_2 at 118 GHz) or if a weaker line (e.g. O_3) happens to lie in the band.

7.2.1 Bandpass measurement

Bandpass calibration usually relies on observing a very strong source for some time; the bandpass functions are obtained by normalizing the observed visibility spectra by their integral over frequency. It is *a priori* not necessary to observe a point source, as long as its visibility can be assumed to be, on all baselines, independent on frequency in the useful bandwidth. If there is some dependence on frequency, then one should take this into account.

7.2.2 IF passband calibration

In many cases the correlator can be split in several independent sub-bands that are centered to different IF frequencies, and thus observe different frequencies in the sky. In principle they can be treated as different receivers since they have different anti-aliasing filters and different delay offsets, due to different lengths of the connecting cables. Thus they need independent bandpass calibrations, which can be done simultaneously on the same strong source.

At millimeter wavelengths strong sources are scarce, and it is more practical to get a relative calibration of the sub-bands by switching the whole IF inputs to a noise source common to all antennas (Fig. 7.1). The switches are inserted before the IFs are split between sub-bands so that the delay offsets of the sub bands are also calibrated out. This has several advantages: the signal to noise ratio observed by observing the noise source is higher than for an astronomical source since it provides fully correlated signals to the correlator; then such a calibration can be done quite often to suppress any gain drift due to thermal variations in the analogue part of the correlator. Since the sensitivity is high, this calibration is done by baseline, so that any closure errors are taken out.

When such an “IF passband calibration” has been applied in real time, only frequency dependent effects occurring in the signal path before the point where the noise source signal is inserted remain to be calibrated. Since at this point the signal is not yet split between sub-bands, the same passband functions are applicable to all correlator sub-bands.

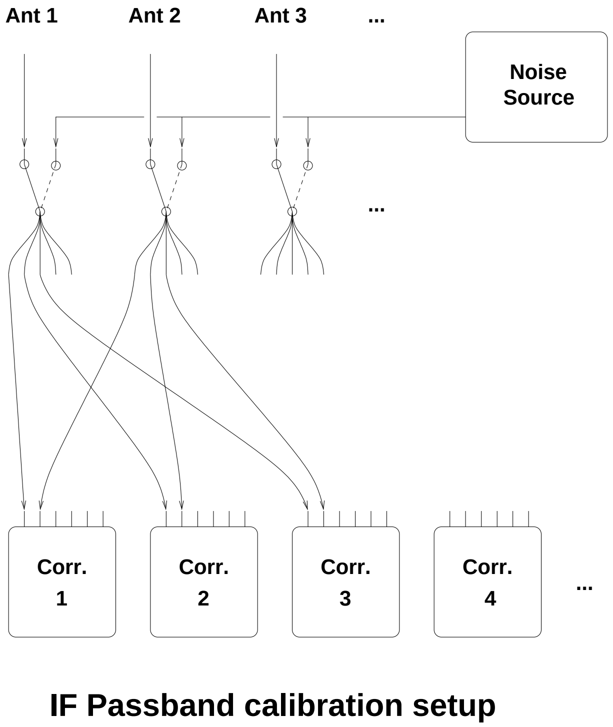


Figure 7.1: IF passband calibration scheme

At Plateau de Bure an "IF passband calibration" is implemented. Of course when the noise source is observed the delay and phase tracking in the last local oscillators (the one in the correlator IF part) are not applied. The precision in phase is $360/\sqrt{\Delta\nu\Delta t} = 0.5°$ at 100 kHz resolution which is sufficient for most projects.

7.2.3 RF bandpass calibration

To actually determine the functions $g_{\mathrm{B}i}(\nu)$ we observe a strong source, with a frequency-independent visibility. The visibilities are

$$\widetilde{V}_{ij}(\nu, t) = g_{\mathrm{B}i}(\nu) g_{\mathrm{B}j}^*(\nu) g_{\mathrm{C}i}(t) g_{\mathrm{C}i}^*(t) V_{ij}(t) \tag{7.14}$$

Then

$$g_{\mathrm{B}i}(\nu) g_{\mathrm{B}j}^*(\nu) = \frac{\widetilde{V}_{ij}(\nu, t)}{\int \widetilde{V}_{ij}(\nu, t) d\nu} \tag{7.15}$$

since the frequency independent factors cancel out in the right-end side. One then averages the measurements on a time long enough to get a decent signal-to-noise ratio. One solves for the antenna-based coefficients in both amplitude and phase; then polynomial amplitude and phase passband curves are fitted to the data.

Applying the passband calibration The passband calibrated visibility data will then be:

$$\widetilde{V}_{\mathrm{C}ij}(\nu, t) = \widetilde{V}_{ij}(\nu, t) / g_{\mathrm{B}i}(\nu) g_{\mathrm{B}j}^*(\nu) \tag{7.16}$$

the amplitude and phase of which should be flat functions of frequency.

Accuracy The most important here is the phase precision: it sets the uncertainty for relative positions of spectral features in the map. A rule of thumb is:

$$\Delta\theta/\theta_B = \Delta\phi/360 \tag{7.17}$$

where θ_B is the synthesized beam, and $\Delta\theta$ the relative position uncertainty. The signal to noise ratio on the bandpass calibration should be better than the signal to noise ration of the spectral features observed; otherwise the relative positional accuracy will be limited by the accuracy of the passband calibration.

The amplitude accuracy can be very important too, for instance when one wants to measure a weak line in front of a strong continuum, in particular for a broad line. In that case one needs to measure the passband with an amplitude accuracy better than that is needed on source to get desired signal to noise ratio. Example: we want to measure a line which is 10% of the continuum, with a SNR of 20 on the line strength; then the SNR on the continuum source should be 200, and the SNR on the passband calibration should be at least as good.

7.2.4 Side band calibration

A millimeter-wave interferometer can be used to record separately the signal in both sidebands of the first LO (see Chapter 5). If the first mixer does not attenuate the image side band, then it is useful to average both side bands for increased continuum sensitivity, both for detecting weaker astronomical sources and increasing the SNR for calibration.

However the relative phases of the two sidebands can be arbitrary (particularly at Plateau de Bure where the IF frequency is variable since the LO2 changes in frequency in parallel with the LO1). This relative phase must be calibrated out. This it is done by measuring the phases of the upper and lower sidebands on the passband calibrator observation. These values can be used later to correct each side band phase to compensate for their phase difference.

During the passband calibration one calculates:

$$e^{i\phi_U} = \frac{\int \widetilde{V}_{ij,USB}(\nu,t)d\nu}{|\int \widetilde{V}_{ij,USB}(\nu,t)d\nu|} \tag{7.18}$$

and

$$e^{i\phi_L} = \frac{\int \widetilde{V}_{ij,LSB}(\nu,t)d\nu}{|\int \widetilde{V}_{ij,LSB}(\nu,t)d\nu|} \tag{7.19}$$

Then at any time the double side band visibility is:

$$\widetilde{V}_{ij,DSB}(\nu,t) = e^{-i\phi_U}\widetilde{V}_{ij,USB}(\nu,t) + e^{-i\phi_L}\widetilde{V}_{ij,LSB}(\nu,t) \tag{7.20}$$

As a result the two terms on the right hand side have zero phase at the time of the pass band calibration and they keep the same phase during the whole observing session.

At observing time, offsets on the first and second LOs can be introduced so that both ϕ_U and ϕ_L are very close to zero when a project is done. This actually done at Plateau de Bure, at the same time when the side band gain ratio is measured (see Chapter 10).

7.3 Phase calibration

We now turn to the more difficult problem of correcting for the time dependence of the complex gains, contained in the functions $g_{\mathrm{C}ij}(t)$. Variations occur in both amplitude and phase. Let us summarize the various effects to be calibrated:

- Interferometer geometry:
  Small errors in the baseline determination will cause slow phase drifts (period 24 hours); unfortunately these errors are dependent on the source direction so they cannot be properly calibrated out by phase referencing on a calibrator, only reduced by a factor of the order of the source to calibrator distance expressed in radians.

- Atmosphere
  The atmosphere introduces phase fluctuations on time scales 1s to a few hours, depending on baseline length and atmospheric conditions (see previous lecture). The effect of fluctuations on short time scales is to cause loss of amplitude by decorrelation, while on the long term the phase fluctuations can be mistaken for structure in the source itself. The critical time there is the time it takes for the projected baseline vector to move by half the diameter of one antenna:
  $$\Delta t_C = \frac{D}{2}\frac{86400}{2\pi b} = 224\frac{D}{15}\frac{450}{b} \tag{7.21}$$
  This time ranges from 4 minutes to 1 hour for Plateau de Bure, depending on the baseline length. The phase fluctuations on short time scales may be corrected by applying the radiometric phase correction method, if it is available.
- Antennas
  The antennas may cause variations in amplitude gains due to degradation in the pointing and in the focusing due to thermal deformations in the antennas. These can be corrected to first order by amplitude calibration but it is much better to keep the errors low by proper frequent monitoring of the pointing and focus, since these errors affect differently sources at the center and at the edges of the field. Focus errors also cause strong phase errors due to the additional path length which is twice the sub reflector motion in the axial direction.
- Electronics
  Phase and amplitude drifts in the electronics are kept low by efficient design (see the lectures on receivers and signal and LO transport). The electronics phase drifts are generally slow and of low amplitude, expect for hardware problems.

A detailed analysis can be found in [Lay 1997].

7.3.1 Phase referencing by a nearby point source

This is the standard, traditional way to calibrate the phases with current interferometers. A point source calibrator is typically observed for $T_{\rm I}$ (a few minutes) every $T_{\rm C}$ (20-30 minutes). One fits a gain curve to the data observed on the calibrator, this gain curve is an estimate of the actual gain curve $g_{{\rm C}i}(t)$. This enables removing most long-term phase drifts from the observation of the target source.

Decorrelation It can be shown that the decorrelation factor for a given baseline is approximated by:

$$f \sim 1 - \frac{1}{2}\int_{-log 2.5 T_{\rm I}}^{\infty} \nu P_\phi(\nu) d(\log\nu) \tag{7.22}$$

The decorrelation is fundamentally a baseline-based quantity: it cannot generally be expressed as a product of antenna-based factors. Both the target source and the calibrator are affected, so amplitude referencing will correct for decorrelation. However, the amount of decorrelation will vary from an integration to the next, so that the amplitude uncertainty is increased.

Phase referencing The slow component of $g_{{\rm C}i}(t)$ to be calibrated out is sampled at intervals $T_{\rm C}$ so that only variations with periods longer than $2T_{\rm C}$ can be followed. However one fits a slow component into the data points so one is sensitive to errors due to the presence of the fast component: the fast component is *aliased* into a slow component. **It is essential to fit a curve that does not go through the points.**

Fast phase referencing One may reduce $T_{\rm C}$ as much as possible to remove a larger part of the atmospheric fluctuation spectrum. Time scales of the order of 10s may be used, at the expense of:

- the time efficiency is decreased (relatively more time is spent on the calibrator resulting in a larger overhead)
- caution must be taken than the time $T_{\rm C}$ may become comparable to the time it takes to water eddies to drift along the apparent distance between source and calibrator.

Water vapor radiometry A radiometry system may be used to monitor the emission of water vapor at a suitable frequency in front of all antennas (dedicated instrument or astronomy receiver; water line or quasi continuum). The fluctuations in the path length can be of a few % of the total path length due to water vapor.

The fluctuations of the water emission are converted into fluctuations of path length by using an atmospheric model (see lecture by M. Bremer, Chapter 9). In principle one could hope to correct for all phase fluctuations this way. However limitations due to receiver stability, to variations in emission from the ground, and to uncertainty in the determination of the emission to path length conversion factor have the consequence that it is not yet possible to consistently correct for the variation of path length between the source and the calibrator.

So far this method at IRAM/Plateau de Bure is used only to correct for on-source fluctuations. Its main effect is to remove the decorrelation effect due to short-term phase fluctuations, improving the precision of amplitude determination.

7.3.2 Phase referencing by a point source in the primary beam

We now consider the simple case where the field contains a strong point source: it can be a continuum source (quasar) or a line source (maser). In that case all phase fluctuations with period longer than $\sim 2T_{\rm I}$ are removed, where $T_{\rm I}$ is the integration time. However statistical errors may be mistaken for true atmospheric phase fluctuations, causing additional decorrelation.

This method gives good results, but for very specific projects which can be observed in very poor atmospheric conditions (e.g. observation of radio emitting quasars, of stars with strong maser lines).

7.3.3 Phase referencing using another band or another frequency

It is generally easier to measure the path lengths fluctuations at a lower frequency (even though the phase scales like frequency), due to both better receiver sensitivity and larger flux of the referencing source. Moreover, in marginal weather conditions, if the rms phase fluctuations at 100 GHz is $\sim 40°$, then at 230 GHz they are of $\sim 100°$, and the phase becomes impossible to track directly due to 2π ambiguities.

If two receivers are available simultaneously, one may subtract to the high frequency phase the phase measured at the low frequency. The atmospheric fluctuations are cancelled and only a slow instrumental drift remains. The gain curve at the high frequency is then determined as the sum of two terms: the low frequency gain curve (including the slow atmospheric terms) plus that slow instrumental drift (which represents any phase fluctuation affecting one of the signal paths of the two receivers).

This method is currently used at Plateau de Bure.

Chapter 8

Atmospheric Absorption

Michel Guélin

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

8.1 The Atmosphere

8.1.1 Constituents of the atmosphere

It is convenient to subdivide the atmosphere into a "clean dry" component, water vapor, and aerosols (water droplets, as well as ice crystals, salt grains & dust particles, which serve as condensation seeds for water).

Abundances Table 8.1 gives the composition of the "clean dry" air in the "standard US" model throughout the troposphere (i.e at altitudes ≤ 20 km). Altogether, N_2, O_2, and Ar represent 99.96 % in volume. Except for CO_2, whose abundance at ground level may vary between day and night by up to a factor of 2, this composition is remarkably homogeneous and constant. It is completed by a number of trace components, most of which are unstable (SO_2, O_3, NO, CO, ...) and whose abundance (in volume) never exceeds 10^{-5}.

The abundance of water is highly variable, but hardly exceed 1% in mass, even locally. Most of the water in the air is in the form of vapor. Even inside the clouds, precipitation and turbulence insure that

Name	Molec. mass amu	Normal abund. (% in volume)	Name	Molec. mass amu	Normal abund. (% in volume)
N_2	28.013	78.084	He	4.003	0.0005
O_2	32.000	20.946	Kr	83.8	0.0001
Ar	39.948	0.934	CH_4	16.043	0.0002
CO_2	44.010	0.033^v	H_2	2.016	0.00005
Ne	20.183	0.0018	N_2O	44.013	0.00005

Table 8.1: Main constituents of the dry air in the troposphere

the mass of water droplets per cm^{-3} seldom equals that of water vapor.

Despite their low abundances, water, which has a large absorption cross section in the IR and a large specific heat of vaporization, ($L_w \sim 600$ cal/g), ozone and carbon dioxyde, which have large IR absorption cross sections, are the major actors of the thermal balance of the air.

Absorption spectrum Ozone is responsible for most of the absorption of the solar radiation in the UV (especially between 1800 and 2900 Å). In the visible, the air is fairly transparent except for scattering by aerosols (mostly water droplets and/or ice crystals when there are clouds). In the infrared, H_2O, CO_2 and, around 10μm, O_3 are very efficient absorbers of the solar and ground radiation.

By clear weather, the atmospheric absorption at mm wavelengths is dominated by rotational and fine structure lines of molecules in their ground electronic and vibrational state. The strongest transitions are electric dipole transitions from polar molecules, like water and ozone. Intrinsically weaker (typically by a factor of 10^{2-3}), but of considerable practical importance, are the magnetic dipole transitions from O_2. When there are clouds, the thin water droplets scatter and absorb the short (mm) radio waves.

8.1.2 Thermodynamics of the air

Ideal gas For one molecule with n degrees of freedom (e.g. in the rigid-molecule approximation, n= 2 (deg. of rotation) +3 (of transl.) = 5 for O_2 and N_2, and n= 3 + 3 = 6 for H_2O), the internal energy per molecule is

$$U = \mathrm{n} \times \frac{1}{2}kT$$

and for $\mathcal{N}$ molecules of O_2 or N_2,

$$U = \frac{5}{2}\mathcal{N}kT = \frac{5}{2}nRT$$

where n is the number of moles and R the gas constant.

The Enthalpy is

$$H = U + pV = U + \mathcal{N}kT$$

$$c_v = \frac{dU}{dT} = \frac{5}{2}\mathcal{N}k\ , \quad c_p = \frac{dH}{dT} = \frac{7}{2}\mathcal{N}k,\ , \quad \gamma = \frac{c_p}{c_v} = \frac{7}{5} = 1.4$$

For adiabatic expansion, we have the standard relations

$$c_p\mathrm{Log}(\frac{T}{T_0}) - R\mathrm{Log}(\frac{p}{p_0}) = 0 \quad \text{and} \quad T = Cst \times p^m$$

where $m = \frac{R}{c_p} = \frac{\gamma-1}{\gamma}$ is the Poisson constant.

Gas mixture: Dalton's law A mixture of ideal gases *i= 1, 2, ...* in a volume V behaves like an ideal gas:

$$\text{Partial pressures :} \quad p_1V = \mathcal{N}_1kT, \quad p_2V = \mathcal{N}_2kT, \quad ... \tag{8.1}$$

$$\text{Total pressure :} \quad pV = (p_1 + p_2)V = (\mathcal{N}_1 + \mathcal{N}_2)kT + ... \quad = \mathcal{N}kT \tag{8.2}$$

Dry air is a mixture of N_2, O_2, ... molecules. It behaves indeed very much like an ideal gas: $R_a = c_{p_a} - c_{v_a} = 8.3143$ J/mol-deg (vs 8.3149 for an ideal gas), γ_a= 1.404 (vs 1.400 for ideal rigid molecules).

Wet air (without clouds) is a mixture of dry air + H_2O molecules. It is customary to denote by e the partial pressure of water vapor, p_a that of dry air, and p' the wet air pressure. The specific heats of water *vapor* are not that different from those of ideal gases: $c_{v_w} = 25.3 + 210^{-3}(T - 273)$; $\gamma_w = 1.37$, vs $c_v = 3R = 24.9$ and $\gamma = \frac{4}{3}$ for a rigid asymmetric top.

Then, Dalton's law yields:

$$c'_v = (1 - \frac{e}{p'})c_{v_a} + \frac{e}{p'}c_{v_w} \simeq (1 + 0.2\frac{e}{p'})c_{v_a} \tag{8.3}$$

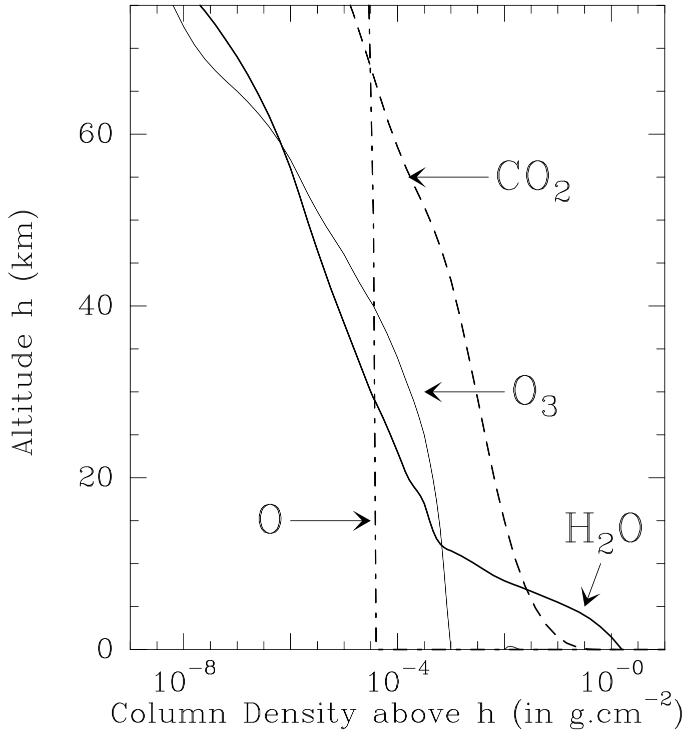


Figure 8.1: Column densities of important atmosphere components above a given altitude – after Queney (1974)

The fractional abundance of water vapor and $\frac{e}{p}$ reaching seldom a few percent, the wet air constants are within a small correction term equal to the dry air constants. In particular, introducing the volume density ρ, it is customary to write:

$$p' = \frac{R\rho' T}{M'} = \frac{R\rho T'}{M_a} \tag{8.4}$$

where $T' = \frac{M'T}{M_a} = T(1 - 0.378\frac{e}{p'})^{-1}$, is the *virtual temperature.*

Then, for the adiabatic expansion of a wet air bubble, one has:

$$T' = Cst \times p^{m'} \tag{8.5}$$

m' is equal to m_a within few per mil, so, in practice, the adiabatic curves of dry air can be used for wet air (without clouds), provided one replaces T by T' (the difference could reach a few K and could be important near 0°C). In the following, we drop the 'prime' signs, except for the virtual temperature T'.

8.1.3 Hydrostatic equilibrium

At large scales, the air pressure and density depend essentially on the massive and slowly varying dry component and are well described by *hydrostatic equilibrium*. The air temperature, as we have seen, depends significantly on the abundance and distribution of water, CO_2 (and O_3 for the stratosphere).

At equilibrium:

$$dp/dh = -g\rho \tag{8.6}$$

$$p = \frac{\rho R T'}{M_a} \tag{8.7}$$

where ρ is the density at an altitude h, p is the pressure, $T' \simeq T$ the air "virtual" temperature. $M_a \simeq 29$ is the average molecular weight, and g the local gravitational field.

$$\frac{dp}{p} = \frac{-gM_a}{RT'}dh \tag{8.8}$$

In the "standard atmosphere" model, T', the temperature of the air varies linearly with altitude and is given throughout the troposphere (i.e. between h=0 and 11 km) by:

$$T' = T'_o - b(h - h_o) \tag{8.9}$$

where $b = 6.5$ K km^{-1} is a constant.

Let us first consider a relatively small change in altitude: $h - h_o \leq 1$ km, $T' \simeq T'_{ave} = (T'(h) + T'(h_o))/2$; we find Laplace's hydrostatic formula:

$$\frac{\rho}{\rho_0} \simeq \frac{p}{p_0} \simeq e^{-\frac{gM_a h}{RT'_{ave}}} = e^{-\frac{h}{h_o}} \tag{8.10}$$

where ρ_o is the density at sea level and $h_o = RT/M_a g = 8.4(T/288)$ km, the scale height. The gas column density (expressed in g.cm^{-2}) along the vertical above a point at sea level is:

$$N_o M_a = \int \rho dh = \rho_o h_o \tag{8.11}$$

and that above a point at an altitude h:

$$N = N_o e^{-h/h_o} \tag{8.12}$$

For larger altitudes, from Eq.8.8 $dh = -dT'/b$, then Eq.8.9 yields

$$\frac{dp}{p} = \frac{-gM_a dh}{bR}\frac{dT'}{T'} \tag{8.13}$$

$$\rho = \rho_o \left(\frac{T'}{T'_o}\right)^{s-1} \tag{8.14}$$

with $s = -g/(bR_a)$.

Although the above equations represent fairly well the density and pressure throughout the troposphere, the temperature distribution can depart significantly from Eq.8.9 near the ground. This latter heats up faster than the transparent air during the day, and cools off more rapidly during the night. The temperature gradient at low altitudes (up to 1-2 km) can be thus steeper or shallower than shown in Eq.8.9. Occasionally, it can be inverted, the temperature increasing with altitude. The *inversion layer* usually stops briskly at 1 or 2 km altitude and the normal temperature gradient resumes above. Inversion layers are common during the night over bare land. They can also be caused by hot winds blowing from the sea.

The local temperature gradient determines stability of the air to vertical motions. A rising bubble of wet air expands adiabatically as long as the water vapor it contains does not condense. Expanding, it cools almost as an ideal gas with:

$$T \propto p^m \text{ and } m = (1 - 0.23q)m_a \simeq m_a \tag{8.15}$$

The pressure is set by the surrounding air (Eq.8.10), and the bubble seen to cool down with an "adiabatic" gradient of

$$\frac{dT}{dh} = -g\frac{M_a}{C_p} = 9.8\text{Kkm}^{-1}$$

If the actual temperature gradient is smaller than the adiabatic gradient, the bubble becomes cooler, hence denser than the surrounding air and its ascent stops. The air is stable. If the local gradient is larger than the adiabatic gradient, the bubble becomes less dense than the surrounding air; the air is unstable and a thick convection layer develops, a situation likely to happen in a hot summer afternoon.

8.1.4 Water

The scale height of water, h_w, which results from a fast evaporation/condensation process, is small (typically 2 km) compared to the equilibrium scale height $h_o = 8.4$ km. At $h = 2.5$ km, the altitude of the Plateau de Bure, the water vapor column density N_w (or w, "amount of precipitable water", when expressed in g.cm^{-2}, or cm of water) is normally reduced by a factor of 3-4, with respect to sea level. This factor, as we have seen, is strongly modulated by the local temperature gradient. w is lower in the presence of a low altitude inversion layer which reduces the vertical turbulence and traps most of the water well below the observatory.

The value of w on a site can be estimated directly by measuring the air pressure p_{tot} (in mbar), temperature T (in K) and relative humidity RH (in percent), and using the formulas 8.7 and 8.11 adapted for water (the air pressure p is replaced by the partial water vapor pressure e):

$$w[\text{mm}] = \rho_w[\text{gm}^{-3}]h_w[\text{km}] \tag{8.16}$$

$$\rho_w[\text{gm}^{-3}] = \frac{eM_w}{RT} = 216.5e[\text{mbar}]/T[\text{K}] \tag{8.17}$$

e is related to $e_s(T)$, the water vapor pressure in saturated air, by:

$$e = RH \cdot e_s/100 \tag{8.18}$$

e_s is a rather complex function of T, for which an approximate analytical expression, as well as tabulated values, can be found in the literature (see e.g. [Queney 1974], p.120). To the zero order, it can be expressed by:

$$e_s[mbar] \simeq 6(T/273)^{18} \tag{8.19}$$

8.1.5 Photodissociation products

Ozone is formed from the reaction of O_2 with atomic oxygen, the latter resulting from the photodissociation of O_2 in the upper atmosphere. Ozone is inexistent below 10 km; its opacity becomes comparable to that of H_2O above 15 km where its relative abundance may reach 10^{-5} (see Fig.8.1). At short millimeter wavelengths, it causes narrow absorption lines (mostly above 230 GHz).

The other trace atmospheric components CH_4, SO_2, CO, N_2O, NO, etc.. have very low abundance and, even polar, play no significant role, their maximum opacities being typically more than 10 times smaller than those of ozone.

8.2 Spectroscopy of H_2O, O_2, O_3

Of the major molecular constituents of the atmosphere (see Table 8.1 and Fig. 8.1), only water vapor and ozone, owing to their bent structure, have a non-zero electric dipole moment. Molecular nitrogen, an homonuclear species, and CO_2, a linear symmetric species, have no permanent electric or magnetic dipole moment in their lowest energy states. These latter molecules, as is the case for 99.9% of gaseous molecules, are singlet states, with electrons arranged two-by-two with opposite spins.

8.2.1 Water vapor

The rotational energy level diagram of water is shown on Fig.8.2. Each level is denoted, as usual for asymmetric top molecules, by three numbers $J_{K_{-1},K_{+1}}$. J, which is a "good" quantum number, represents the total angular momentum of the molecule; by analogy with symmetric tops, K_{-1} and K_{+1} stand for the rotational angular momenta around the axis of least and greatest inertia. Allowed radiative transitions obey the selection rules $\Delta J = \pm 1, \Delta K = \pm 1, 3$, with K_{-1}, K_{+1} : $odd, odd \leftrightarrow even, even$ or $o, e \leftrightarrow e, o$. The levels with K_{-1} and K_{+1} of the same parity are called *para* levels, those of opposite parity, *ortho* levels. Transitions between ortho and para levels are forbidden; due to the presence of two symmetrical hydrogen

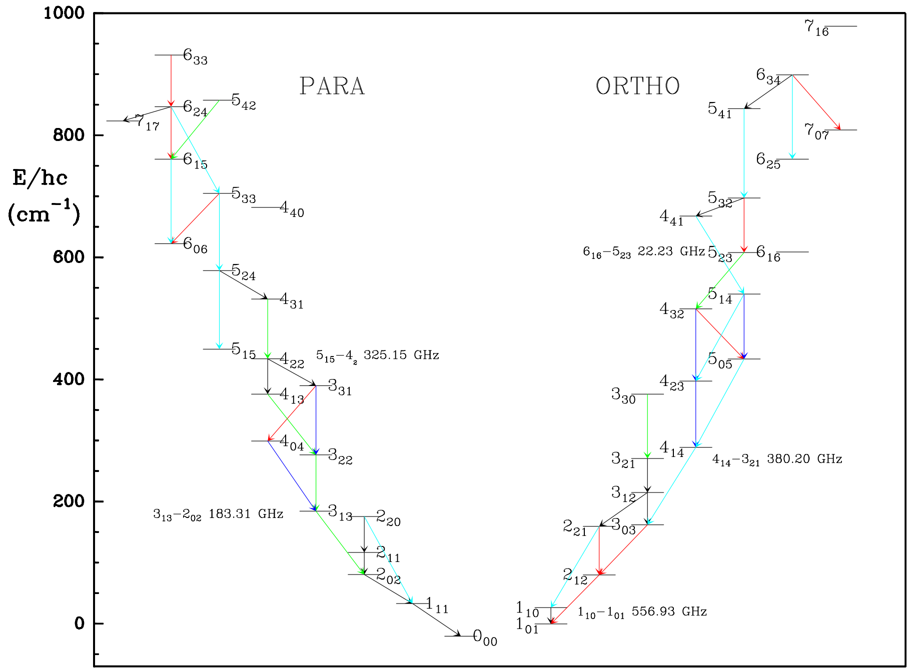


Figure 8.2: The rotational energy level diagram of water vapor

nuclei, the ortho levels have a nuclear statistical weight 3 times larger than the para levels of same J (see e.g. [Townes & Schawlow 1975]). The arrows in Fig.8.2 show the transitions which will be observed by the satellite FIRST (J.Cernicharo, private communication).

8.2.2 Molecular oxygen

Molecular oxygen, although homonuclear, hence with zero electric dipole moment, has a triplet electronic ground state, with two electrons paired with parallel spins. The resulting electronic spin couples efficiently with the magnetic fields caused by the end-over-end rotation of the molecule, yielding a "large" magnetic dipole moment, $\mu^{mag} = 10^{-20}$ emu. The magnetic dipole transitions of O_2 have intrinsic strengths $\sim 10^{2-3}$ times weaker than the water transitions. O_2, however, is 10^{2-3} times more abundant than H_2O, so that the atmospheric lines of the two species have comparable intensities.

The spin of 1 makes of the ground electronic state of O_2 a triplet state ($^3\Sigma$). **N**, the rotational angular momentum couples with **S**, the electronic spin, to give **J** the total angular momentum: $\mathbf{N} + \mathbf{S} = \mathbf{J}$. The $\mathbf{N}\cdot\mathbf{S}$ interaction (and the electronic angular momentum–electronic spin interaction $\mathbf{L}\cdot\mathbf{S}$) split each rotational level of rotational quantum number $N \geq 1$ into three sublevels with total quantum numbers

$$J = N+1, J = N \quad \text{and} \quad J = N-1$$

the $J = N+1$ and $J = N-1$ sublevels lying below the $J = N$ sublevel by approximately $119(N+1)/(2N+3)$ GHz and $119/(2N-1)$ GHz, respectively ([Townes & Schawlow 1975] p.182). Note that the two identical ^{16}O nuclei have spins equal to zero and obey the Bose-Einstein statistics; there are only odd N rotational levels in such a molecule.

The magnetic dipole transitions obey the rules $\Delta N = 0, \pm 2$ and $\Delta J = 0, \pm 1$. Transitions within the fine structure sublevels of a rotational level (i.e. $\Delta N = 0$) are thus allowed. The first such transition is the $(J, N) = 1, 1 \leftarrow 0, 1$ transition, which has a frequency of 118.75 GHz. The second, the $1, 1 \leftarrow 2, 1$ transition, has a frequency of 56.26 GHz. It is surrounded by a forest of other fine structure transitions with frequencies ranging from 53 GHz to 66 GHz. The first "true" rotational transition, the $N = 3 \leftarrow 1$ transitions, have frequencies above 368 GHz (368.5, 424.8, and 487.3 GHz).

The rare isotopomer $^{18}O^{16}O$ is not homonuclear, hence has odd N levels and a non-zero electric dipole moment. This latter, however, is vanishingly small (10^{-5}D). $^{18}O^{16}O$, moreover, has a very low abundance

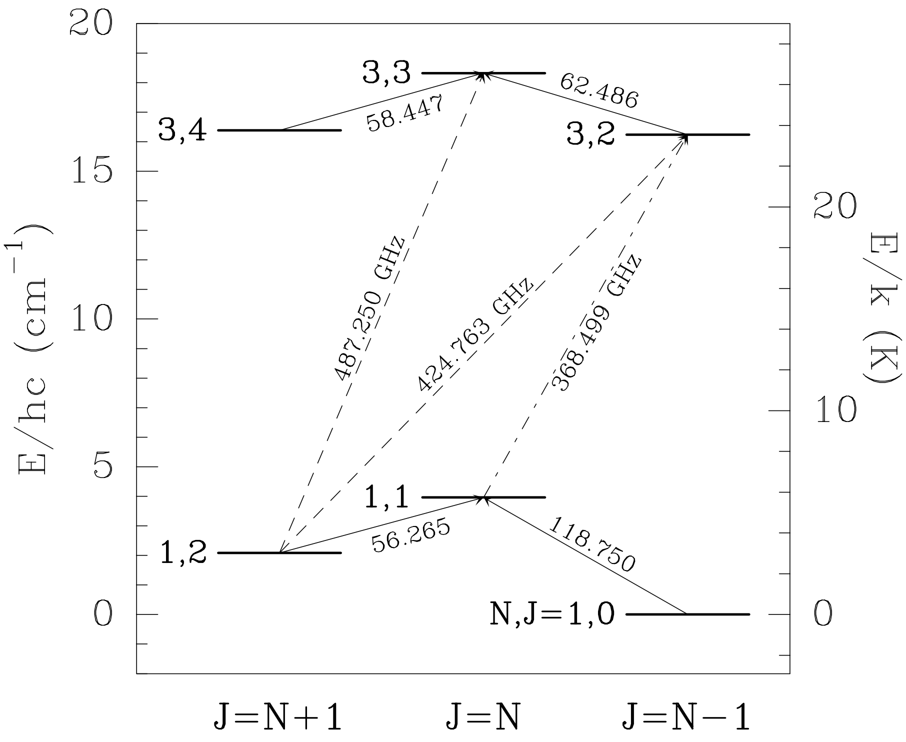


Figure 8.3: The rotational energy level diagram of molecular oxygen

(few hundred times smaller than the main isotopomer), so that its magnetic dipole transitions (even the $\Delta N = 2$, which have stronger intrinsic strengths), can be neglected.

The line opacity and absorption coefficients of $^{16}O_2$ are given by relations similar to Eq.8.24 and Eq.8.25.

8.2.3 Ozone

As noted above, ozone is mostly concentrated between 11 and 40 km altitude; it shows large seasonal and, mostly, latitude variation. Because of its high altitude location, its lines are narrow: at 25 km, ρ_a, hence $\Delta\nu$, is reduced by a factor of 20 with respect to see level; moreover, the dipole moment of ozone (μ= 0.53 D), 3.5 times smaller than that of H_2O, further reduces the ozone line widths.

Because of their small widths and despite the small ozone abundance, ozone lines have significant peak opacities, especially above 230 GHz ($\tau_o = 0.2 - 0.3$).

8.3 Propagation of a wave in the atmosphere – Line shapes

A plane wave propagating along the z direction in the atmosphere can be represented by its electric field vector $\mathbf{E}$:

$$\mathbf{E}(z,t) = \mathbf{E}_m e^{-j2\pi\nu(t - n_c \frac{z}{c})} \tag{8.20}$$

where $n_c = n + jn_i$ is the complex refraction index.

The real part, the true refraction index, $n = c/v_p$ (where v_p is the phase velocity of the wave), is often expressed in terms of the refractivity,

$$N = 10^6(n-1) \simeq 0.223\rho_a + 1760\rho_w/T \tag{8.21}$$

The right hand sum, which was empirically derived, is known as the Smith-Weintraub equation. It separates the contribution of the dry air component (first term, where ρ_a is the air density, expressed in gm^{-3}) from that of water vapor (second term, where ρ_w is the density of water vapor).

The imaginary part of the refraction index, n_j causes an exponential attenuation of the wave amplitude, and is related to the power absorption coefficient κ_ν by:

$$n_j = \frac{c\kappa_\nu}{4\pi\nu} \tag{8.22}$$

At a frequency ν close to a rotational transition $\nu_{lu} = (E_u - E_l)/h$, the optical depth is:

$$\tau_\nu = \int \kappa_\nu(T, P)dh \tag{8.23}$$

κ_{lu}, the absorption coefficient integrated over the transition lu is given by the standard asymmetric top formula (see e.g. [Townes & Schawlow 1975] p.102):

$$\kappa_{lu} = \frac{8\pi^2 h^{1.5}}{3c(kT)^{2.5}}(\frac{\rho_w}{m_w})g_I\sqrt{ABC}\mu^2 S_{lu}T^{-2.5}(1 - \frac{h\nu}{kT})e^{E_l/kT}\nu_{lu}^2\Phi(\nu - \nu_o) \tag{8.24}$$

Replacing A, B, C, the rotational constants (here in Hz), and $\mu = 1.85$ Debye $= 1.85\,10^{-18}$ esu.cm, the electric dipole moment, by their values for H_2O, and setting $(1 - \frac{h\nu}{kT}) = 1$,

$$\kappa_{lu}^w[cm^{-1}] = 5.710^{-24}(\rho_w/m_w)g_I S_{lu}(T/273)^{-2.5}e^{E_l/kT}\nu_{lu}^2\Phi(\nu - \nu_o) \tag{8.25}$$

S_{lu} is the transition intrinsic strength, g_I= 3/2 or 1/2 is the nuclear relative statistical weight of the ortho and para levels (see below), and $\Phi(\nu - \nu_o)$ the line profile. ν_{lu} is now in GHz.

The line profile is given to a good approximation by the well known Van Vleck and Weisskopf collisional profile (here, multiplied by π, [Townes & Schawlow 1975] p.342):

$$\Phi(\nu - \nu_o) = \frac{\Delta\nu}{(\nu - \nu_o)^2 + (\Delta\nu)^2} + \frac{\Delta\nu}{(\nu + \nu_o)^2 + (\Delta\nu)^2} \tag{8.26}$$

with $\Delta\nu = \frac{1}{2\pi\tau}$, where $\Delta\nu$ is the line width and τ the mean time between molecular collisions.

At the centre of the line, the second term of Eq.8.26 becomes negligible, and

$$\Phi(\nu - \nu_o) = 1/\Delta\nu = 2\pi\tau \sim \frac{2\pi}{\frac{\rho_a}{m_a}v\sigma^2} \tag{8.27}$$

$v \sim \sqrt{T}$ is the molecular velocity and σ the collisional cross section.

Note that the density in Eq.8.24 is ρ_w, whereas that in Eq.8.27 is ρ_a: the absorption coefficient at the centre of the line is proportional to ρ_w/ρ_a. It is independent of the total air pressure, as long as this ratio (hence RH) stays constant. This is not the case, of course, away from the centre, since $\Delta\nu \sim \rho$: as the density drops, the lines become narrower and narrower. In the far wings of the line, the second term in Eq.8.26 and the contribution from the wings of other lines cannot anymore be neglected. In fact, a better fit to the water emission in the far wings is reached if Eq.8.26 is replaced by another collisional line shape, called the "kinetic profile"

$$\Phi(\nu - \nu_0) = \frac{4\nu\nu_0\Delta\nu}{(\nu^2 - \nu_0^2)^2 + 4\nu^2\Delta\nu^2} \tag{8.28}$$

The opacity and width of the main $H_2^{16}O$ lines are large: $\tau_o = 60$ and $\Delta\nu \simeq 20$GHz, for the fundamental line in normal conditions of p and T, and for $\rho_w = 1$g.m^{-3} and w_w= 1 mm (dry weather). These lines dominate most of the millimetre and submillimetre atmospheric attenuation; deviations from theoretical line shapes in the far wings (typically 1/10th of intensity) are accounted for by an empirical continuum. The rare isotopomer $H_2^{18}O$, a few hundred times less abundant than $H_2^{16}O$, makes a negligible contribution (see the discussion by [Waters 1976])

8.4 The atmospheric absorption spectrum at millimeter wavelengths, ATM

Calculations of zenith atmosphere opacity at 2.5 and 2.9 km, the altitude of the IRAM sites, have been made by [Cernicharo 1988] and [Cernicharo & Pardo 1999]. A computer programme, ATM, repeating these calculations has been installed on-line on the IRAM telescopes of Pico Veleta and Plateau de Bure; it is activated at each calibration or skydip and allows to interpret the observed sky emissivity in terms of water and oxygen contributions and of upper and lower sideband opacities. (During skydips, the antenna is pointed successively at different elevations and the emission of the sky measured at each step; the sky emission variation is fitted by an exponential function of the air mass, and the atmosphere opacity and average temperature readily derived). Note that the opacities derived from sky emissivity observations do not always agree with those calculated from the measurement of p, T, and RH on the site, as water vapor is not at hydrostatic equilibrium.

Some of the results for the band 20–500 GHz are shown on Fig.8.4. One recognizes from left to right, the (blended) forest of fine structure transitions from O_2, near 60 GHz, the $1,1 \leftarrow 0,1$ fine structure line of O_2 at 118.75 GHz, the third lowest lines of para water (still 200 K above the ground level), at 183.31 GHz, and the fourth ortho water line (420 K above ortho ground level), at 325.15 GHz. The fundamental line of ortho water ($1_{10} \leftarrow 1_{01}$), at 556.94 GHz is visible. The water and oxygen lines delineate the 4 atmospheric "windows" of the millimetre spectrum (called the 3 mm, 2 mm, 1.3 mm and 0.8 mm windows). Water is seen to dominate completely atmospheric absorption above 150 GHz.

Less than one millimeter of precipitable water vapor corresponds to exceptionally good winter weather conditions on the sites of Pico Veleta and Plateau de Bure. Such conditions seldom happen even on Mauna Kea. Two to three millimeter of water are standard by clear winter nights at these observatories; six to ten millimeter of water are typical of clear summer nights.

The typical zenith atmosphere opacities, in the dips of the 1.3 mm and 0.8 mm windows (e.g. at the frequencies of the $J = 2-1$ and $3-2$ rotational transitions of CO, 230.54 and 345.80 GHz), are respectively 0.15–0.2 and 0.5–0.7 in winter. The astronomical signals at these frequencies are attenuated by factors of respectively $\simeq 1.2$ and 2 at zenith, 1.3 and 2.8 at 45 degree elevation, and 1.7 and 6 at 20 degree elevation. Larger attenuations are the rule in summer and in winter by less favorable conditions.

The $J = 1-0$ line of CO, at 115.27 GHz, is close to the 118.75 GHz oxygen line. Although this latter is relatively narrow, it raises by $\simeq$0.3 the atmosphere opacity (which is 0.35–0.4). The atmosphere attenuation is then intermediate between those at 230 and 345 GHz (by dry weather, however, it is more stable than the latter, since the water contribution is small). The measurement of accurate CO line intensity ratios (even not considering the problems linked to differences in beam size and receiver sideband gain ratios) requires therefore good weather, a high source elevation, and a careful monitoring of the atmosphere.

A catalogue of lines intensities in several standard astronomical sources, measured with the IRAM 30 m telescope has been published [Mauersberger et al 1989]. The lines intensities were calibrated by the chopper-wheel method, following the above recipes. The reader is referred to this report for details.

8.5 Correction for atmospheric absorption, T_A^*

By analogy with the Rayleigh-Jeans approximation, $I = 2kT/\lambda^2$, which strictly applies to long wavelengths, the mm-wave radio astronomers have introduced the concept of "radiation" or "effective" temperatures, which scale *linearly* with the detected power.

The noise power detected by the telescope is the sum of the power received by the antenna, $\mathcal{W}_A$, and of the noise generated by the receiver and transmission lines, $\mathcal{W}_{rec}$.

Using Nyquist's relation $\mathcal{W} = kT\Delta\nu$, $\mathcal{W}_A$ and $\mathcal{W}_{rec}$ can be expressed in terms of the temperatures T_A and T_{rec} of two resistors, located at the end of the transmission line, which would yield noise powers equal to $\mathcal{W}_A$ and $\mathcal{W}_{rec}$, respectively:

$$\mathcal{W}_A + \mathcal{W}_{rec} = kT_A\Delta\nu + kT_{rec}\Delta\nu = k(T_A + T_{rec})\Delta\nu.$$

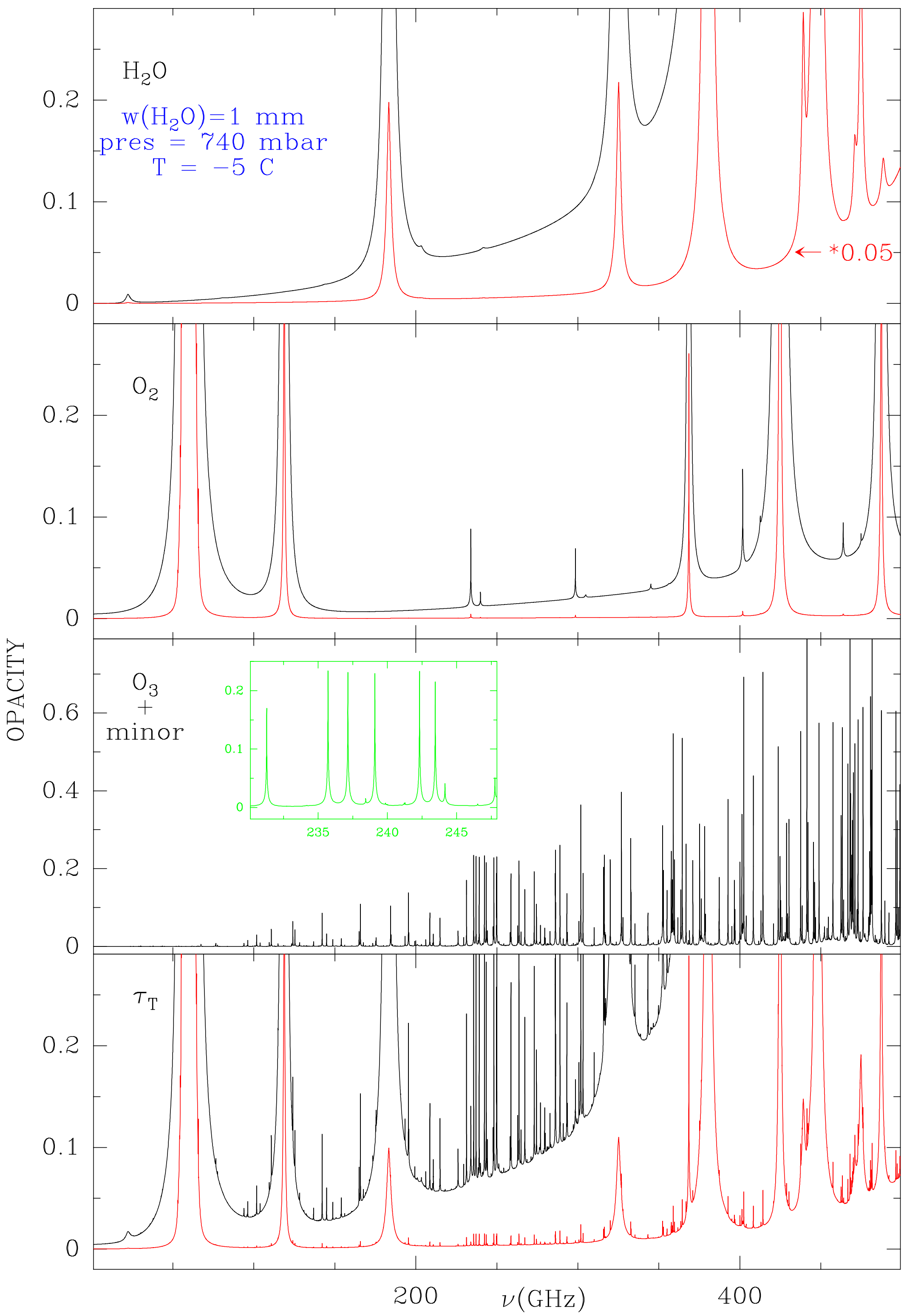


Figure 8.4: Zenith opacity of the standard winter atmosphere at an altitude of 2.5 km for 1 mm of precipitable water vapor, as a function of frequency (GHz). The contributions of H_2O, O_2 and O_3 are shown in the three upper panels. The red line shows the same zenith opacity at a scale of 1/20. One may note the importance of the water line wings above 150 GHz, compared to those of O_2 (a consequence of the absence of electric dipole moment in the latter molecule) and O_3. Courtesy [Cernicharo & Pardo 1999]

T_A is called the "antenna temperature" and T_{rec} the "receiver temperature". T_A becomes T_{load} when the receiver horn sees a load, instead of the antenna, and T_{gr} when it sees the ground. It should be noted that T_{load} and T_{gr} are not stricto sensus equal to the load and ground physical temperatures, but are only "Rayleigh-Jeans" equivalent of these temperatures (they are proportional to the radiated power). For ambient loads, they approach however closely the physical temperature, since $h\nu/k \simeq 11$ K at $\lambda = 1.3$ mm.

When observing with the antenna a source and an adjacent emission-free reference field, one sees a change $\Delta T_A = T_A(sou) - T_A(ref)$ in antenna temperature.

Because of the calibration method explained below, it is customary, in mm-wave astronomy, to replace ΔT_A, the source antenna temperature, by ΔT_A^*, the source antenna temperature corrected for atmosphere absorption and spillover losses. Both are related through:

$$T_A = (1 - \eta_f)T_{gr} + \eta_f(T_{sky} + \Delta T_A^* e^{-\tau})$$

where τ is the line-of-sight atmosphere opacity. η_f and η_b are the forward and beam efficiency factors, which denote the fractions of the power radiated by the antenna on the sky and in the main beam, respectively (they are typically of the order of 0.9 and 0.7).

The source equivalent "radiation temperature" T_R (often improperly called "brightness temperature" and therefore denoted T_{MB} when it is averaged over the main beam) and ΔT_A^* are related through

$$\Delta T_A^* = \int_{sou} T_R \mathcal{A}(x,y)dxdy$$

where $\mathcal{A}(x,y)$ is the antenna power pattern. For a source smaller than the main beam, $\Delta T_A * = \eta_b T_{MB}$.

When observing a small astronomical source with an antenna temperature $\Delta T'_A >> T_{BG} = 2.7$ K, located at an elevation el, one detects a signal $\mathcal{V}$ (of scale: $\mathcal{G}$ volt or counts per Kelvin):

$$\frac{\mathcal{V}_{sour}}{\mathcal{G}} = M_{sour} = T_{rec} + (1-\eta_f)T_{gr} + \eta_f T_{sky} + \eta_b \Delta T'_A e^{-\tau} \tag{8.29}$$

This signal can be compared with the signals observed on the blank sky (T_{atm}), close to the source, and to that observed on a hot load (T_{load}):

$$\begin{aligned} M_{atm} &= T_{rec} + (1-\eta_f)T_{gr} + \eta_f T_{sky} \qquad (8.30) \\ T_{sky} &= (1 - e^{-\tau})T_{atm} \\ M_{load} &= T_{rec} + T_{load} \end{aligned}$$

here, we have neglected the cosmologic background and assume, in a first step, that the receiver is tuned single sideband.

8.5.1 Simplest case

Let's assume that $T_{load} \simeq T_{gr} \simeq T_{atm}$; then:

$$\begin{aligned} M_{load} - M_{atm} &= \eta_f T_{gr} e^{-\tau} \\ M_{sou} - M_{atm} &= \Delta T_A^* e^{-\tau} \\ \Delta T_A^* &= \frac{M_{sou} - M_{atm}}{M_{load} - M_{atm}} \eta_f T_{gr} \qquad (8.31) \end{aligned}$$

Note that in Eq.8.31, the measurement of the antenna temperature includes the atmospheric opacity correction, but does not depend explicitly on an assumption on the atmospheric opacity. We can write:

$$\Delta T_A^* = \frac{M_{sou} - M_{atm}}{M_{load} - M_{atm}} \mathbf{T_{cal}} \tag{8.32}$$

where we define T_{cal} as $T_{cal} = \eta_f T_{gr} \simeq \eta_f T_{atm}$.

8.5.2 More realistic case

Typically, the mean atmosphere temperature is lower than the ambient temperature near the ground by about 40 K: $T_{atm} \simeq T_{gr} - 40$ K; then, the formula above still holds *if we replace* T_{cal} by:

$$\begin{aligned} T_{cal} &= (T_{load} - T_{emi})e^{\tau} && (8.33) \\ \text{with } T_{emi} &= T_{sky}\eta_f + (1-\eta_f)T_{gr} \\ &= \frac{(T_{load} + T_{rec}) \times M_{atm}}{M_{load}} - T_{rec} && (8.34) \\ T_{sky} &= (1 - e^{-\tau})(T_{gr} - 40) \end{aligned}$$

T_{rec}, the receiver effective temperature is usually calculated by the Y factor method using a cold load (usually cooled in liquid nitrogen, i.e. at 77 K) and an ambient load (e.g. at 290 K).

$$\begin{aligned} Y &= \frac{M_{hot\_load}}{M_{cold\_load}} \\ T_{rec} &= \frac{T_{hot\_load} - YT_{cold\_load}}{Y-1} && (8.35) \end{aligned}$$

8.5.3 General case

The receiver is not purely single-sideband. Let us denote by G^l and G^u the normalized gains in the receiver lower and upper sidebands, $G^l + G^u = 1$. The atmosphere opacity per km varies with altitude as does the air temperature.

Then, the above expressions of T_{sky} and T_{emi} should be explicited for each sideband ($j = u$ or l):

$$\begin{aligned} T^j_{sky} &= (1 - e^{-\tau^j})T^j_{atm} && (8.36) \\ T_{emi} &= T^l_{sky}\eta_f G^l + T^u_{sky}\eta_f G^u + (1-\eta_f)T_{gr} && (8.37) \end{aligned}$$

The atmospheric transmission model ATM [Cernicharo 1988] allows to calculate iteratively τ_ν from a load+ sky measurement. The values of τ^l, τ^u are calculated for the Standard US atmosphere (parameters are: Winter-, Spring-, or Summer-temperature T, altitude, latitude, water vapor w) by summing up the contributions of O_2 ($A_\nu(T)$), H_2O ($B_\nu(T)$) and O_3 ($C_\nu(T)$) (including rare isotopes and vibrationally excited states):

$$\tau_\nu = A_\nu(T) + wB_\nu(T) + C_\nu(T) \qquad (8.38)$$

Depending on ν, the values of $A_\nu(T), B_\nu(T), C_\nu(T)$ are derived either from a Table or estimated from empirical formulae. A first guess of the amount of precipitable water is made from the ambient temperature, pressure and humidity by using the relations of Sec.8.1.4. Then, the expected T_{sky} and T_{emi} are calculated from the two expressions above and T_{emi} is compared to its value measured from the the observation of the atmosphere and the load (Eq.8.35). The value of w is changed and the calculation of τ^j, T^j_{sky} and T_{emi} restarted. Normally, the process converges after 2 to 4 iterations. Once w and T_{emi} are known, the calibration factor T_{cal} can be derived

$$T^j_{cal} = (T_{load} - T_{emi})e^{\tau^j} \qquad (8.39)$$

and the data calibrated in the T^*_A scale using Eq.8.32.

Chapter 9

Atmospheric Fluctuations

Michael Bremer

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

9.1 Introduction

We have already encountered the effects of atmospheric absorption in the lecture by Michel Guélin (Chapter 8). Even under low opacity conditions, observations can be difficult or impossible due to atmospheric phase fluctuations. In principle, integration of an interferometric signal is like the adding up of vectors. The amplitude is the length of a vector, and its orientation is given by the phase. Errors in the phase will cause part of the amplitudes to cancel each other out according to Eq. 9.1

$$\overline{V_{ij}^{m}} = V_{ij} \exp\left(-\sigma_{\phi}^{2}/2\right) \tag{9.1}$$

where V_{ij} are the ideal visibilities, $\overline{V_{ij}^{m}}$ the integrated ones, and σ_{ϕ} the phase noise in radian (assuming a Gaussian noise distribution).

If our eyes were sensitive in the millimeter range with a resolving power of some arc seconds, we would not only see a luminous sky where sources are difficult to make out on the background: we would notice refracting bubbles of sizes between some centimeters to several kilometers drifting with the wind, merging, separating and distorting the view behind them. It is water vapor which has not attained the concentrations necessary for cloud formation, and mixes badly with dry air. Even in the case of low opacity and a sky clear for human perception, these distortions - which are shifts and tilts in the incoming phase front - can be so important that they make observations impossible.

In this lecture we will study the physics behind this effect and possible ways to correct for it which have become available in recent years.

The phases of a wave front that reaches ground-based observatories have been modified by their journey through refractive index variations in the Earth's atmosphere. The instrument itself has no way to tell which part of the phases are due to valuable structure in the astronomical source and which part was caused during the last kilometers in front of the telescope.

As a result, the apparent position of a point source keeps moving, so that details of an extended source become blurred ("seeing"). The perturbations become progressively decorrelated with increasing separation between two lines of sight.

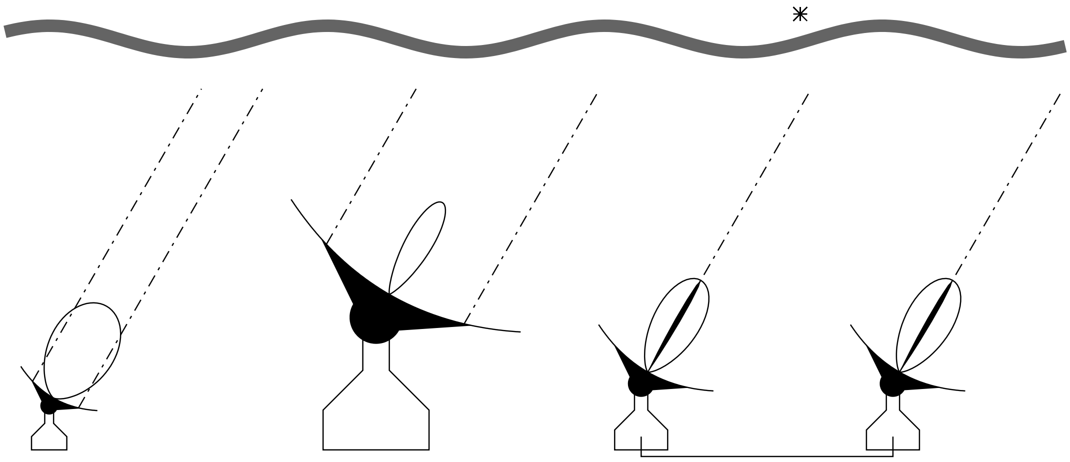


Figure 9.1: Two single dish radio telescopes and one synthesis array observing a source. Angular resolution and the effects of atmospheric turbulence increase with the diameter of an instrument. An interferometer's synthesized beam has the same problems that a single dish with a projected baseline's diameter would experience.

Depending on turbulence scale, wavelength and telescope size, a ground-based observer will be confronted with different effects. To some extent, one can reduce unwanted perturbations by choosing an observatory's site carefully. But there are always some days which are better than others.

- If the telescope diameter is much larger than a typical turbulence cell and if phase shifts over a turbulent cell are less than 1 radian, one works in the **diffractive** regime: some power is scattered into an error beam, but the diffraction limited resolution of the instrument is conserved. These conditions can be found for very long baseline interferometry at cm wavelengths under good conditions.
- In the **refractive** regime, the turbulence cells are much larger than the telescope so that the whole image seems to move around. Phase shifts can be several radians. *This is the case we have to cope with at the Plateau de Bure interferometer.*
- In the **intermediate** case, one gets speckles on short integration times which average into an image convolved with a seeing disk (an effect well-known to optical and near infrared observers).

For a small radio telescope, the phase noise passes unnoticed (Fig.9.1). Its beam size is bigger than the apparent position shifts induced by seeing, and the ray paths leading to the opposing outer edges of the reflector will not differ much. For a big single dish, the effect can become noticeable: under unstable conditions, a source may wander in and out of the beam, disturbing not only the observations of the target of interest but also pointing and focus. An interferometer suffers even more from seeing because the distances between individual antennas are large. This has the double effect of making the synthesized beam smaller and of increasing the differences in the atmospheric turbulence pattern between the optical paths. In the optical and near infrared, it is the dry atmosphere which causes the phase shifts. In the centimeter radio range, it is the tropospheric water vapor and the ionospheric plasma irregularities, and mainly water vapor in the millimeter radio range.

The interest of interferometric phase correction is to determine and remove the atmospheric phase noise. Each individual antenna has to correct a time variable "piston" in the incoming wave front, which is the equivalent of adaptive optics for the synthesized dish.

Most of the water vapor is confined to the troposphere (<10 km) with an exponential scale height of 2 km. Its molecular weight is 18.2 g/mol so that it has the tendency to rise above dry air (28.96 g/mol). Our planet retains the water vapor due to the negative vertical temperature gradient in the troposphere, and the fact that H_2O is close to the transition points to liquid and solid under terrestrial pressure and

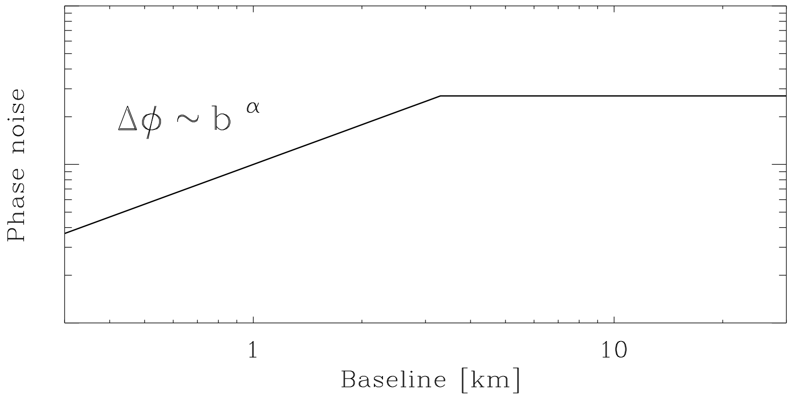


Figure 9.2: Atmospheric phase noise as a function of baseline length. The position of the break and the maximum phase noise are weather dependent.

temperature conditions. Venus is an example for a planet with a hot troposphere (extreme CO_2 and H_2O greenhouse effect) who lost most of its water long ago.

Under Earth's environmental conditions, water vapor mixes badly with dry air and tends to form bubbles with sizes up to several kilometers, which are broken up by turbulent motions. When choosing high and dry sites for millimeter astronomy, one must keep in mind that high phase noise can be induced even by low amounts of precipitable water.

Parameters that influence interferometric phase noise strongly are:

- wind speed,
- projected baseline length,
- the time of the day (nights are typically more stable due to the missing solar energy input to turbulence),
- topographic effects, e.g. the turbulent wakes of mountains.

The leveling off of the phase noise on a high, constant value corresponds to the outer scale size of the turbulence, where the fluctuations along two lines of sight become totally decorrelated, i.e. as bad as they can possibly get (Fig. 9.2). Increasing the baselines will not increase the phase noise any more, and this is why very long baseline interferometry can work.

In the following, some fundamentals about turbulence will be explained, and how one can analyze the properties of a turbulent atmosphere. Possible methods to monitor the phase fluctuations will be discussed, with emphasis on the phase monitoring system on the Plateau de Bure. Some examples with observational data will be presented, demonstrating its benefits and current limitations.

9.2 Hydrodynamical basics of turbulent motion

Turbulence has its origin in the non-linearities of the hydrodynamical equations of motion. These equations are not derived from first principles, but present the easiest consistent way of describing the motion of compressible (gases) or incompressible (liquids) media. Given the complexity of observed hydrodynamical phenomena (white water rivers, curling smoke, ...) it was doubted for more than a century if the equations could really be as simple as given below.

Scepticism was nourished by the fact that some stationary (i.e. time invariant) solutions were mathematically valid but not observed experimentally, which can be understood today by checking the solutions for their stability. It is an interesting fact that stationary solutions are *not* always the most stable ones

– this assumption might appear "natural", but is in fact quite misleading. To find out under which conditions one has to expect turbulence, we must have a closer look into the equations of hydrodynamics.

In a free flow (i.e. no outer confinements like tubes) and under sub-sonic conditions, one can treat air as an incompressible medium. This allows to use the Navier-Stokes equations (Navier 1827, extended by Stokes in 1845) which describe such a medium with viscosity (Eq. 9.2).

$$\underbrace{\frac{d\vec{V}}{dt}}_{1.} + \underbrace{\frac{grad\,P}{\rho}}_{2.} + \underbrace{grad\,U}_{3.} - \underbrace{\frac{\eta}{\rho}}_{4.}\Delta\vec{V} = 0 \tag{9.2}$$

The respective terms describe:

1. This part is the complete derivative in time, containing non-linear terms:

$$\frac{d\vec{V}}{dt} = \frac{\delta\vec{V}}{\delta t} + \left(\vec{V}grad\right)\vec{V} = \frac{\delta\vec{V}}{\delta t} + \underbrace{\frac{1}{2}\,grad\left(\vec{V}\cdot\vec{V}\right)}_{kin.\ Energy} - \underbrace{\vec{V}\times\left(\vec{\nabla}\times\vec{V}\right)}_{Vortices} \tag{9.3}$$

2. The pressure term describes the reaction to external forces, and is related to density and temperature through the equation of state.

3. External forces, described by the gradient of a Potential U. This form makes it easier to include this term into the others. We are interested here in cases where U is constant in good approximation, so we can leave this term out in the following.

4. Energy dissipation. Viscous terms, with the Laplacian operator $\Delta = \nabla^2 = div\ grad$. The material constant η is the viscosity coefficient for incompressive media (there is a term with a second coefficient for compressive media, but it can be neglected here).

To solve Equation 9.2, one needs:

- the equation of state for the gas,
- the conservation of mass and energy,
- the boundary conditions.

We will not have to deal with these equations directly. If you have to solve them for some reason: There is a whole class of numerical methods and library codes in the literature, which avoid non-evident pitfalls concerning different coordinate systems and numerical stability. For the current discussion, we are only interested in the question of similarity.

Flow problems resemble each other for certain combinations of flow velocity, spatial dimension and viscosity. This allows to predict the general properties of a hydrodynamic system (and if it is turbulent or not) from small models or other well-studied cases where the geometries are the same. It is convenient to change to dimensionless equations by expressing lengths and velocities in units of the system's typical length scale l_0 (e.g. the size of an obstacle) and the unperturbated flow velocity V_0.

One obtains not only $V' = V/V_0$ and $l' = l/l_0$, but must change all other units which are combinations of the two: $t' = t \cdot v_0/l_0$, $\rho' = \rho \cdot l_0^3$ and $P' = P \cdot l_0^3/v_0^2$. As a result, we get the dimensionless equation Eq.9.4:

$$\frac{dV'}{dt'} + \frac{grad\,P'}{\rho'} - \frac{1}{Re}\Delta V' = 0 \tag{9.4}$$

which contains **Reynold's number** $Re = l_0 \cdot V_0 \cdot \rho/\eta$. It determines the relative influence of the energy dissipating term relative to the non-linear turbulent term. A high Reynold's number will reduce the effect of $\Delta V'$, so that turbulence will develop. Each problem has its specific **Critical Reynold's number** Re_c,

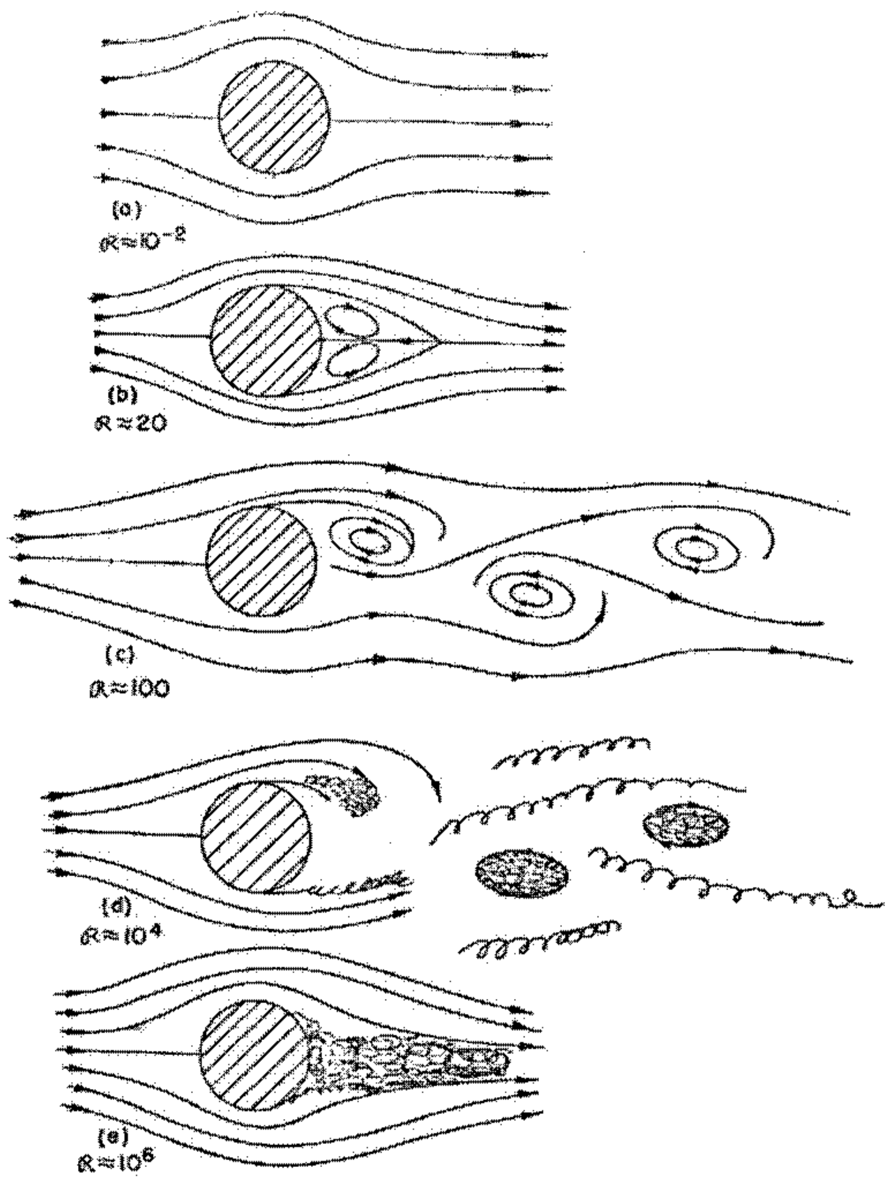


Figure 9.3: Flow around a cylindric obstacle for various Reynold's numbers (from Feynman).

Medium	η [g/cm s]	$\nu = \eta/\rho$ [cm^2/s]
Water	0.010	0.010
Air	0.00018	0.150
Alcohol	0.018	0.022
Glycerin	8.5	6.8
Mercury	0.0156	0.0012

Table 9.1: Examples for viscosity and the kinetic viscosity η/ρ various media at 20°C (from Landau-Lifschitz). The viscosity changes little with altitude, so that the density dependency of ν dominates and satellites above ≈ 300 km altitude experience a non-turbulent, laminar flow.

which lies typically between 10 and 100. For a given geometry and a $Re > Re_c$, no stable solution exists (Fig.9.3).

With Eq.9.4 and the viscosities from Tab.9.1, we find that for ambient conditions, wind moving faster than 1 cm/s hitting an obstacle bigger than 1 cm generates a turbulent flow. A house (10 m) in a 10 m/s wind has $Re \approx 6 \cdot 10^6$, mountains $Re \approx 1 \cdot 10^9$.

This means that turbulence is something quite common in our environment. In our daily life, we encounter many turbulent systems which defy detailed prediction: leaves of wind-moved trees, eddies in flowing water, the structure of clouds, to name only a few. All these effects are non-linear and will never repeat themselves exactly, although their parameters stay within certain limits.

It may be surprising that these phenomena have been neglected by classical physics for centuries, to become finally popular in the wake of chaos theory and the development of powerful computers. One reason for this was surely the problem of repeating an experiment with inherent chaos exactly, and the sheer bulk of work for doing the calculations. Linear physics were favored not because they are more abundant in nature, but because they were easier to understand and reproduce.

Even chaos theory has a hard time with the atmosphere. The famous “strange attractors” which describe non-repetitive curves in solution spaces are not very useful for turbulence, because one is not sure if the number of solution space dimensions is finite or not in this case.

Meteorology is a well-known (and sometimes notorious) example for predictions of a non-linear system. In spite of a network of measurement stations and satellite data, boundary conditions are not known precisely enough for long term forecasts. This is the famous “butterfly effect”: A butterfly moving its wings in South America may change the weather in Europe six months later.

But don’t start hunting butterflys to prevent storms right now: this example only illustrates that non-linear systems do not obey the “small cause - small effect” rule of linear physics but a rather imprecise “small cause at the right place may have a big effect” rule. So it could be a butterfly, a spoken word, a thought, or a tree falling in a forest, or all of them together that may tip a balance.

For all practical applications, the interactions are too complex to backtrack the cause – but nevertheless, the history that makes a leaf move right now in a unique way outside your window contains somewhere the gravitational pull of faraway galaxies, and your mere presence will leave traces in the turbulent structures all over the universe (no liabilities implied, fortunately).

A small perturbation may set off a non-linear cause-and-effect chain, but this is only because a turbulent system can have multiple macroscopically different states without violating the conservation of energy. Using a statistical approach, we will now discuss turbulence’s energy distribution on different scale sizes. This is an important step to understand how atmospheric parameters (including phase noise) change with time and distance.

9.3 Statistical properties of turbulence

We start with a simple model of turbulence. It must explain why the scale size of the finest turbulence structures becomes smaller and smaller with increasing *Re* (see Fig.9.3), and should allow to treat the finest details in a homogeneous way. It cannot explain why certain structures form and not others, but it describes the average flow of energy across the scale sizes of turbulence.

- Kinetic energy enters the medium on large scales, in the form of convection or friction on an obstacle (**energy range**).
- The energy is transferred towards smaller scale sizes over eddy fragmentation, while the Reynolds number decreases (**inertia range**).
- The smallest eddies have sub-critical Re's, dissipate heat, and are stable (**viscous range**).

[Kolmogorov 1941] advanced a hypothesis for high ($Re > 10^6 - 10^7$) Reynold's numbers, postulating that turbulence in the inertia range was determined only by one parameter ϵ (kinetic energy converted to heat by viscous friction per unit time and unit mass). In the viscous range, it would only depend on ϵ and the already discussed viscosity η. This model treats cases like the seemingly amorph eddies-within-eddies part in Fig.9.3 (d),(e). As we have derived in the previous section, (d) and (e) are indeed the cases to be expected in the troposphere.

The inertia range is interesting for us because it corresponds to spatial dimensions of some meters to 2-3 kilometers, i.e. the baselines of the PdBI fall into this range.

For the mathematical treatment of highly developed turbulence, one can use a formalism based on random variations.

Most "classical" statistics represent a given distribution of probability (binomial, Poisson, Gaussian, ...) around a most likely measurement value. For atmospheric parameters like e.g. temperature and wind velocity, we must make a more general approach: the most likely measurement values vary with time and space, which means they can be represented by *non-stationary random processes*. The classical average and its variation are not very useful to describe these systems.

An instrument for the characterization of non-stationary random variables are *structure functions*, which were first introduced by [Kolmogorov 1941]. A scalar structure function has the form given in Eq.9.5,

$$D_f(x_i, x_j) = \overline{(f(x_i) - f(x_j))^2} \tag{9.5}$$

i.e. a function $f(x)$ is measured at the positions x_i and x_j, squared and averaged over many samples to obtain a $D_f(x_i, x_j)$. When the average level of f changes, the average differences between $f(x_i)$ and $f(x_j)$ stay constant.

The structure function formalism can even be used to describe vector parameters like the turbulent wind velocity, in this case one simply needs 3×3 tensors of structure functions for their description. We won't need tensors in the following discussion, however. The detailed mathematical formalism of random fields would be too much for this course (see [Tatarski 1971] for details). We will only discuss the basic concepts and their application to phase shifts.

Real atmospheric parameters are functions of time and space. For time dependency, Taylor's hypothesis of frozen turbulence has been quite successful (Eq.9.6). It states that the pattern of refractive index variations stays fixed while it is moving with the wind.

$$f(x, t + t') = f(x - V_{al} \cdot t', t) \tag{9.6}$$

This means that for the structure functions, one can either measure at two different sites simultaneously or measure in one place and compare different times. Time-like structure functions are often easier to determine because the sampling is continuous and instrumental effects are reduced by averaging. The velocity V_{al} is also called "Velocity aloft" and can differ notably from measurements of a ground-based meteorological station: wind speeds increase with altitude and change direction due to the diminishing effect of ground friction.

For two measurement points which are a distance r apart from each other, one finds that the structure functions of many atmospheric parameters (temperature, refractive index, absolute wind velocity, ...) obey a $r^{2/3}$ power law. This law can be derived from the theory of random fields, but the easiest way is as follows:

Consider a velocity fluctuation δV_r (where δV_r may be large) which occurs on a scale size r and a time $t = r/\delta V_r$:

- its energy per mass unit is $\propto (\delta V_r)^2$

- The energy per mass and time: $\epsilon \propto \delta V_r^2/t = \delta V_r^3/r$
- Stationary transport of this energy from large to small scale sizes, where it is finally dissipated
- Approximately, $D_{rr}(r) = \overline{(V(r_l + r) - V(r_l))^2}$ is dominated by eddies of size r, i.e. $D_{rr}(r) \approx (\delta V_r)^2$
- Therefore, the formula for intermediate scale sizes is:

$$D_{rr}(r) \propto (\epsilon \cdot r)^{2/3} \tag{9.7}$$

For a thin layer, refractive index fluctuations and phase fluctuations are identical. This is the *thin screen approximation.*

$$D_{\varphi}(r) = C \cdot (\epsilon \cdot r)^{2/3} \tag{9.8}$$

In a thick turbulent layer, the phase front encounters multiple refractive index fluctuations and the power law index changes. This problem can be solved by analyzing the irregular refracting medium over its Fourier transform. After [Tatarski 1961], the spectral density of the function

$$D(r) = r^p \tag{9.9}$$

with $0 < p < 2$ is in the three dimensional case

$$F'(\kappa) = \frac{\Gamma(p+2)}{4\pi^2} sin(\pi p/2) \kappa^{-(p+3)} \tag{9.10}$$

An important condition is that the fluctuations must have an outer limit, i.e. that the power law does not increase indefinitely. To get the phase fluctuations from the refractive index spectrum, we take

$$D_{\varphi}(r) = 4\pi \int_0^{\infty} [1 - J_0(\kappa r)] F'(\kappa) \kappa d\kappa \tag{9.11}$$

with the Bessel function J_0 and finally obtain the power law for thick screen turbulence:

$$D_{\varphi}(r) \propto (\epsilon \cdot r)^{5/3} \tag{9.12}$$

For the phase noise ($\Delta\varphi(r) = \sqrt{D_{\varphi}(r)}$), one has therefore to expect power laws with exponents between 1/3 and 5/6. The absolute scaling factor for the power law and the position of the break where the phase noise levels off depend on the observing site, and of course on the weather.

Due to the quasi-random nature of phase fluctuations, forecasts and inter/extrapolations can be considered inadequate for a phase correction system. Direct measurements of the water vapor column along the line of sight are therefore the most reliable approach.

9.4 Remote sounding techniques

These methods were originally developed for meteorology and control of industrial emissions.

1. **LIDAR**: compares laser back scattering or transmission. DIAL: differential absorption LIDAR, works at two frequencies (on and off the atmospheric line of interest), detects 0.01 g/m^3 water vapor. Disadvantage: works best from aircraft, expensive equipment.
2. **SODAR**: Remote sounding with sound waves. Detects turbulence, but gives little quantitative results.
3. **IR window**: H_2O line absorption in front of a strong continuum source (Sun, Moon, Jupiter). Disadvantage: Directions of observing and monitoring beam differ. The phase correction degrades as a function of the separation angle and the distance of the dominant turbulent layer.

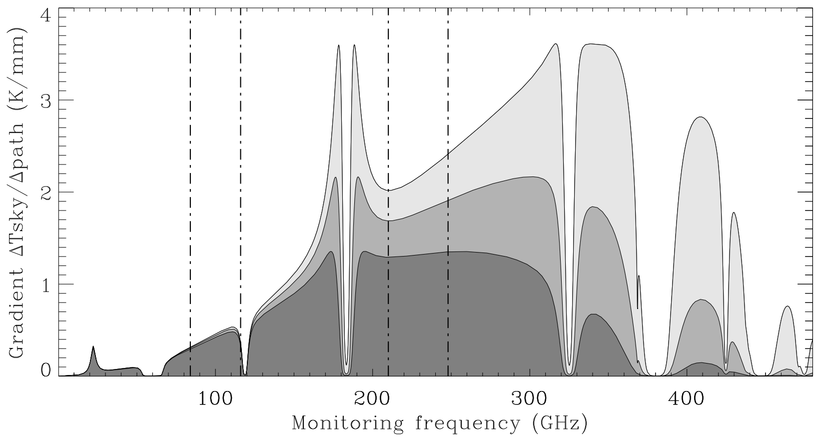


Figure 9.4: Gradient $\Delta T_{sky}/\Delta path$ (K/mm) as a function of frequency and total precipitable water under clear sky conditions. The atmospheric model assumed an ambient temperature of 275 K, pressure 780 mbar, elevation 45°, an observing frequency of 90 GHz and various amounts of water vapor. Light grey: 3 mm water, middle grey: 5 mm water, dark grey: 8 mm water. Dash-dotted lines indicate the receiver tuning ranges of the PdBI.

4. **Radiometric**: Uses the atmospheric emission. Dedicated monitors operate mostly near the 22 GHz or 183 GHz lines (several spectral channels). The inter-line regions of the 1mm and 3mm windows are also sensitive enough, but make it difficult to remove cloud emission.

For the radiometric approach, it is useful to study the sensitivity as a function of frequency, i.e. by how much the sky emission changes for a fixed fluctuation of water vapor, which corresponds to a fixed wet path fluctuation. Fig.9.4 shows what change in T_{sky} one must measure under conditions of various humidity.

There are two reasons to use the 22.2 GHz line: Clouds are easier corrected at this frequency, and receiver components are less expensive.

One notices that the 84-116 GHz window is 1-2 times as sensitive as the 22.2 GHz line, and the 210-248 GHz window 4.5-8.3 times. A dedicated receiver near the 183 GHz water line would have the highest sensitivity, but can suffer from temperature dependent saturation effects. It is better adapted for sites where the total amount of water above the instrument is typically less than 3 mm.

9.5 Current phase correction at IRAM

Remote sounding is done with the astronomical 1mm receivers in the inter-line region at the chosen observing frequency. One uses the total power channel (bandwidth 500 MHz). Advantages of this approach are the close coincidence of observed and monitored line of sight, and the fact that no additional monitoring equipment is needed.

First success was on April 18, 1995, with the installation of the present receiver generation on the PdBI [Bremer 1995]. Critical advantages were the improved total power stability of the receivers and the capability to observe in the 1 mm window. The necessary stability for a 30° phase rms at 230 GHz is about $\Delta M/M = 2 \cdot 10^{-4}$.

Steps of the method:

1. Calibration of the total power counts M_{atm} to T_{sky} as given in the lecture on amplitude and flux calibration.

2. Iterate the amount of precipitable water vapor in an atmospheric model to reproduce T_{sky}. There is no “learning phase” of the algorithm on a quasar, just the model prediction.

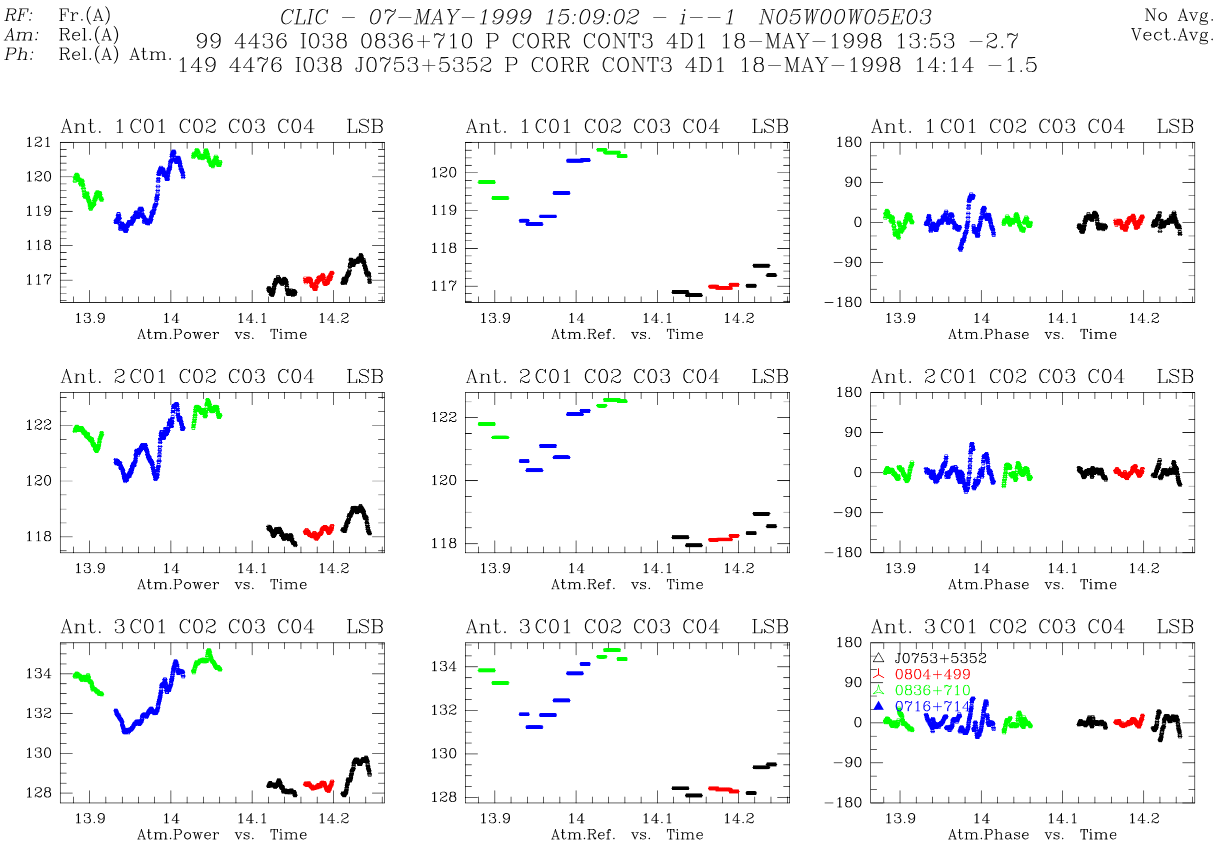


Figure 9.5: Antenna based total power at 228.3 GHz, the reference value to calculate the differential correction, and the model-based phase shift per antenna at 86.2 GHz.

3. The amount of water vapor along the line of sight is proportional to the wet path length.

$$path \approx 6.7 \cdot water(\text{Zenith}) \cdot airmass(\text{Elevation}) \tag{9.13}$$

However, wet path length and opacity have different dependencies on frequency, atmospheric pressure and temperature which should be taken into account. The main increase of the refractive index n of water vapor relative to dry air happens in the infrared, which makes it difficult to use the Cramers-Kronig relations linking it to opacity (integration over many transitions). For simplicity, we use the calculations by [Hill & Cliffort 1981] for the frequency dependency and the temperature and pressure dependencies by [Thayer 1974] instead. These references use not n but the refractivity N, which is defined over the excess path length L relative to vacuum propagation over the line of sight s:

$$L = 10^{-6} \int N_\nu(s) ds \tag{9.14}$$

$$N(P,T) = 77.493 \frac{p_{atm}}{T} - 12.8 \frac{p_V}{T} + 3.776 \times \frac{p_V}{T^2} \tag{9.15}$$

Hill and Cliffort calculate $N(\nu)$ for $T = 300\,\text{K}$, $P = 1013.3\,\text{mbar}$, 80% humidity

$$N(p,T,\nu) = 77.493 \frac{p_{atm}}{T} - 12.8 \frac{p_V}{T} + N(\nu) \frac{p_V}{28.2} \left(\frac{300.}{T} \right)^2 \tag{9.16}$$

4. Subtract the average over a time interval (default: the duration of a scan) to remove residual offsets due to receiver drift and ground pickup, which can be different for each antenna (see Fig.9.5).

5. Convert the antenna specific path shifts into phase at the observed wavelength, $\Delta\phi_i$

6. Calculate the baseline specific phase shifts $\Delta\phi_{ij} = \Delta\phi_i - \Delta\phi_j$. A corrected and an uncorrected version are calculated and stored during the real time reduction which compresses the spectra over one scan. The precision of the correction in relative pathlength is about 65μm per antenna (hence 90μm per baseline, i.e. $\sqrt{2}$ larger).

7. During the off-line data reduction, the user can choose freely between the corrected and uncorrected sets. The phase correction can fail under the following conditions:

   - Clouds: the model only works for clear sky conditions, and will over-estimate the phase shifts seriously in the presence of clouds.
   - Very stable winter conditions: The phase noise of the observations can be below 25° at 230 GHz, which is the intrinsic noise of the correction method.
   - Total power instabilities: For some frequencies, the receivers are difficult to tune. One can get a nice gain in the interferometric amplitude, but an unstable total power signal with an intrinsic noise well above 25° at 230 GHz.

   Even for the cases above, the observer has lost nothing because the uncorrected scans are still there. Software tools are available which help to decide when to apply the correction.

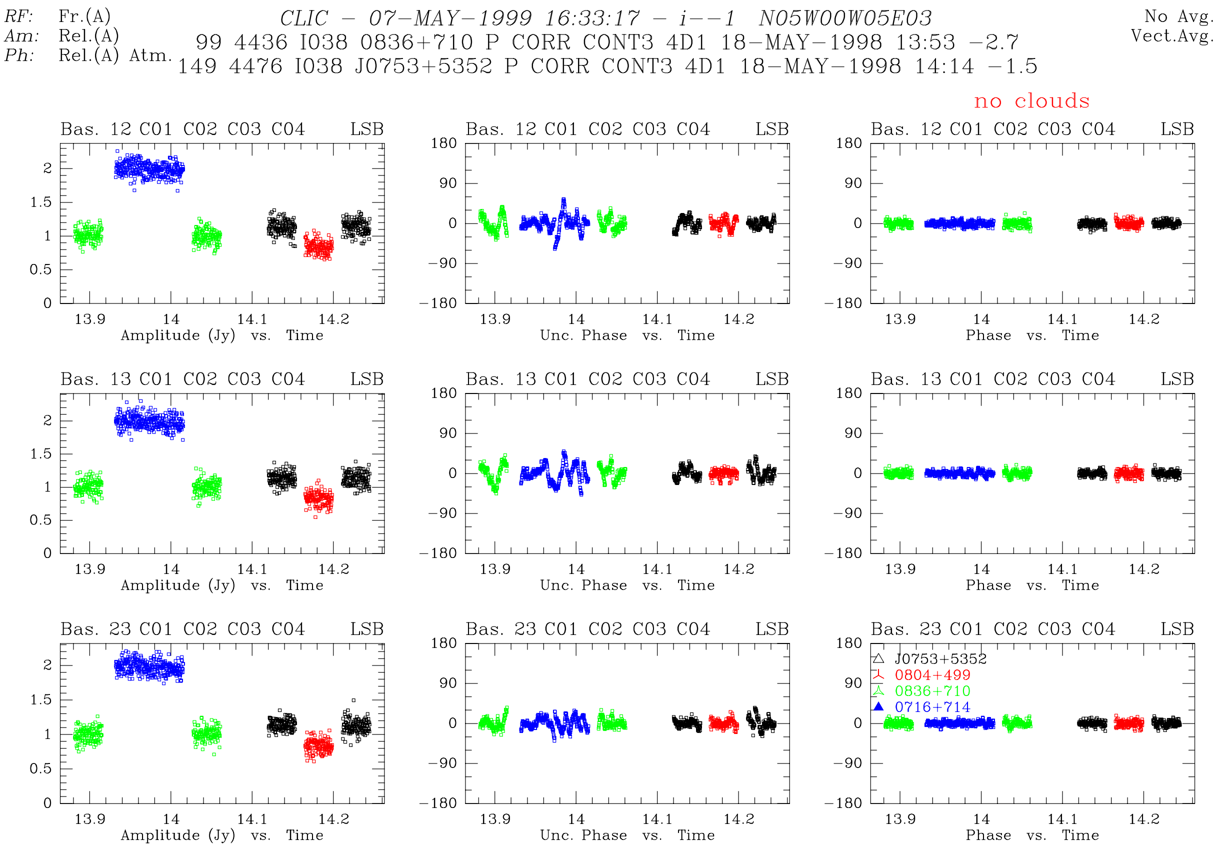


Figure 9.6: Baseline based amplitudes, uncorrected phase and monitor corrected phase at 86.2 GHz with a time resolution of 1 s. The data correspond to the antenna based section in Fig.9.5. The phase calibration applied in columns 2 and 3 was obtained using `STORE PHASE /SELF` on a one minute time scale, thereby setting the mean phases to zero.

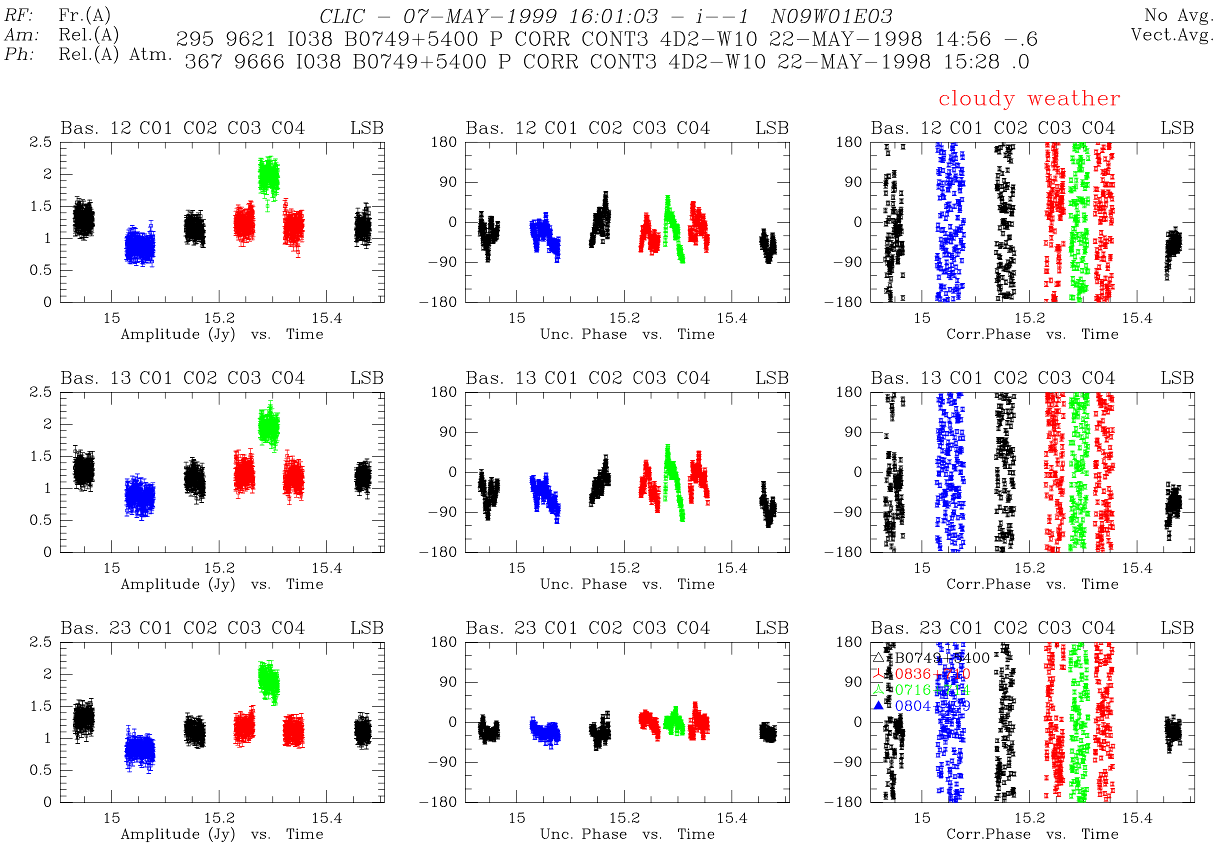


Figure 9.7: Baseline based amplitudes, uncorrected phase and monitor corrected phase at 86.2 GHz for a cloudy data section. The clear sky model over-estimates the correction and would result in serious amplitude loss. The off-line data reduction software will identify those parts and disable the correction there.

9.6 Phase correction during off-line data reduction

In a typical data reduction session, the atmospheric phase correction hides behind two unobtrusive buttons (Fig.9.8). The default of the `PHCOR` button will calculate the phase correction on a scan basis, i.e. only the corrected amplitudes will be used and not the phases. There are two reasons to be cautious about the corrected phases:

- As we have learned from turbulence theory, averages are *not* good in describing data which obey a structure function. Adjusting the monitored phases to zero average can introduce random-like offsets due to slow (large-scale) components of the atmospheric fluctuations. You can only rely on averages if the outer scale of turbulence have passed several times over the instrument. This can be true on your source (typically 20 min integration time), but is doubtful on the calibrators (integrations of typically 3 min). It may work for compact configurations depending on wind speed, i.e. it depends on the weather. There may be an improvement, but it cannot be guaranteed.

- Nearly linear changes in total power can be due to a big water vapor bubble in the atmosphere or a gain drift in the receivers. The first will produce a phase shift, the second will not, and the software cannot tell them apart. In most cases, gain drifts happen when the antenna has just moved a large distance in elevation, as pump friction and liquid helium distribution change with the receiver cabin tilt. Such drifts are invisible in the interferometric amplitude (opacity correction), but the phase correction is more sensitive to them.

The part of related header variables is accessible over
`CLIC> VARIABLES MONITOR`
Some items are scan-based in time resolution, others can be expanded to a time scale of one second over the total power signal. One can plot the following quantities related to the phase correction:

- `ATM_POWER`: total power counts of the monitoring receiver
- `ATM_REFERENCE`: the offsets which are subtracted from `ATM_POWER`
- `ATM_EMISSION`: the calibrated sky emission in Kelvin
- `ATM_PHASE`: the modeled atmospheric phase
- `ATM_UNCORRPH`: the astronomical phase uncorrected from atmosphere
- `ATM_CORRPH`: the astronomical phase corrected with the model
- `ATM_VALIDITY`: 0 or 1 whether the phase correction has been declared valid or not. This flag is antenna specific.

In order to check the validity of the phase correction, the standard reduction runs the command
`CLIC> STORE CORRECTION AUTO 15`
i.e. CLIC will test the phase calibrator observations (type P) if the application of the phase correction improves or degrades the amplitudes, and will declare the correction on source observations (type O) in a ±15 minutes time window for good.

Apart from `AUTO`, one can use `GOOD` and `BAD` for a manual override of the diagnostics, and `SELF` to check the amplitude for each scan (indifferent to type O or P) for strong sources.

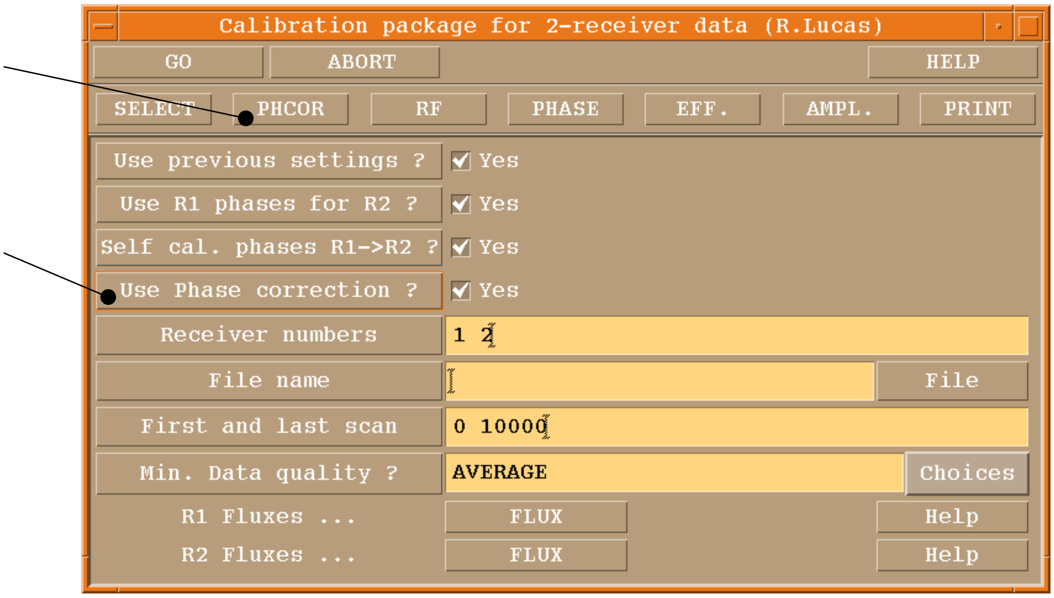


Figure 9.8: The CLIC calibration package menu, with phase correction related items marked.

9.7 Frequently asked questions

- **How often is the phase correction applied?** Statistics have been calculated for all receiver 1 CORR scans in the preliminary data reduction files between September 1997 and March 1999 (a total of 3120 hours). Over this period, the phase correction was diagnosed "good" during 78.8% of the observing time.

- **Does one gain something on a self calibrated source?** Yes, the monitor corrected spectra are improved on a sub-scan scale, i.e. on one second. An example for this case can be found in Fig. 9.6.
- **How fast can the phase change?** Phases at 3 mm can turn by more than 360° over 30 seconds. Even under calm conditions, it was found that a time resolution of one second for the correction gave better results than an average correction over four seconds.
- **Is dry weather the same as low phase noise weather?** Unfortunately not. Even with 3mm precipitable water, one can have very bad phases if the wind is high (about 10 m/s).
- **Why is the phase correction sometimes disabled under clear sky conditions?** Each correction system has its intrinsic noise. If the atmospheric phase noise is below 10° at 3 mm or 30° at 230 GHz (which happens under stable winter conditions), the added noise will undo the benefits, and the diagnostics will switch the correction off. The same can happen under less favorable conditions if a receiver has an unstable total power signal (this can happen at some frequencies).
- **How does the corrected phase depend on baseline length?** Basically, it becomes independent. Over time scales longer than the monitor time interval, the structure function dependence stays.
- **Are there still changes / improvements in the system?** Yes, we are working on it.

Chapter 10

Amplitude and Flux Calibration

Anne Dutrey

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

Calibration compensates for imperfections and unknowns in the instrument use, including antenna defects (surface quality, focus), pointing errors, atmospheric transmission and fluctuations, receiver and backend gain and instabilities, etc.... In addition, calibrated data are also expressed in a reliable physical unit.

This lecture is then cut in three parts, of equal importance:

1. The single-dish calibration of the amplitude: its errors and biases
2. The flux density calibration
3. The final amplitude calibration of interferometric data

10.1 Definition and Formalism

From R.Lucas lecture 7, Eq. 7.1), the baseline-based observed visibility $\widetilde{V}_{ij}(t)$ is linked to the true visibility V_{ij} of the source by:

$$\widetilde{V}_{ij}(t) = \mathcal{G}_{ij} V_{ij} + \epsilon_{ij}(t) + \eta_{ij}(t) \tag{10.1}$$

In antenna-based calibration, $\mathcal{G}_{ij}$ can also be written as:

$$\mathcal{G}_{ij} = g_i(t) g_j^*(t) = a_i(t) a_j(t) e^{i(\phi_i(t) - \phi_j(t))} \tag{10.2}$$

Therefore, for antenna i, the antenna-based amplitude correction for the Lower sideband a_i^L is given by

$$a_i^L(t) = T_{cal\,i}^L(t) G_i^L(\nu, t) \mathcal{B}_i(t) \tag{10.3}$$

and for the Upper sideband:

$$a_i^U(t) = T_{cal\,i}^U(t) G_i^U(\nu, t) \mathcal{B}_i(t) \tag{10.4}$$

where $T^U_{cal_i}$ and $T^L_{cal_i}$ are the corrections for the atmospheric absorption (see lecture 8 by M.Guélin), in the Upper and Lower sidebands respectively. $\mathcal{B}_i$ the antenna gain (affected by pointing errors, defocusing, surface status and systematic elevation effects). Note that Eqs.10.3-10.4 do not include the decorrelation factor f (see lecture 7 by R.Lucas) because this parameter is baseline-based. We assume here decorrelation is small enough, i.e. $f = 1$; if not, a baseline-based amplitude calibration may be required.

$G^L_i(\nu,t)$ and $G^U_i(\nu,t)$ are the electronics gains (IF chain+receiver) in the Lower and Upper sidebands, respectively. The receiver sideband gain ratio is defined as $G^{UL}_i(\nu,t) = G^U_i(\nu,t)/G^L_i(\nu,t)$. The sideband gain ratio is to first order independent of the frequency ν within the IF bandwidth. The derivation of the receiver gains is given in Chapter 7. At Bure, the receivers and the IF chain are very stable and these values are constant with time (and equal to G^{UL}_i, G^U_i and G^L_i, respectively, since we also neglected their frequency dependence). They are measured at the beginning of each project on a strong astronomical source. Moreover in Eq.10.3-10.4, we use the fact that for a given tuning, only the receiver gains and the atmospheric absorption have a significant dependence as a function of frequency.

Section 10.2 will focus on the corrections for the atmospheric absorption ($T^U_{cal_i}(t)$,$T^L_{cal_i}(t)$) and the possible biases they can introduce in the amplitude.

In the equations above, the amplitudes can be expressed either in Kelvin (antenna temperature scale, T^*_A, $\eta_B = \eta_F$) or in Jy (flux density unit, 1 Jy= 10^{-26} $\mathrm{W\,m^{-2}Hz^{-1}}$). The derivation of the conversion factor between Jy and K, in Jy/K, $\mathcal{J}_{iS}$ (single-dish mode) and $\mathcal{J}_{iI}$ (interferometric mode) and its biases will be detailed in section 10.3 which is devoted to the flux density calibration.

Finally Section 10.4 will deal with the understanding of the terms $\mathcal{B}_i(t)$ and f, the amplitude calibration of interferometric data.

10.2 Single-dish Calibration of the Amplitude

This calibration is done automatically and in *real-time* but it can be redone *a posteriori* if one or several parameters are wrong using the CLIC command `ATMOSPHERE`. However, for 99 % of the projects, the single-dish calibration is correct. Moreover, we will see in this section that in most cases, even with erroneous calibration parameters, it is almost impossible to do an error larger than $\sim 5\%$.

For details about the properties of the atmosphere, the reader has to refer to the lecture by M.Bremer (Chapter 9) while the transmission of the atmosphere at mm wavelengths is described in the lecture of M.Guélin (Chapter 8). Most of this lecture is extracted from the documentation "Amplitude Calibration" by S.Guilloteau for single-dish telescope and from [Guilloteau et al 1993].

Since all this part of the calibration is purely antenna dependent and in order to simplify the equations, the subscript $_i$ will be systematically ignored. In the same spirit, the equations will be expressed in T^*_A scale taking $\eta_f = \eta_B$.

The atmospheric absorption (*e.g.* for the Lower side-band T^L_{cal}) can be expressed by

$$T^L_{cal} = \frac{(T_{load}(1+G^{UL}) - T^L_{emi} - G^{UL}T^U_{emi})}{e^{-\tau^L/\sin(Elevation)}} \tag{10.5}$$

where T_{load} is the hot load and T^L_{emi} and T^U_{emi} are the noise temperature received from the sky in the lower an upper sidebands respectively (for the IRAM interferometer, the difference in frequency between the upper and lower sidebands is ~ 3 GHz).

The system temperature T_{sys} is given by:

$$T^L_{sys} = T^L_{cal} \times \frac{M_{atm}}{M_{load} - M_{cold}} \tag{10.6}$$

The main goal of the single-dish calibration is to measure T_{cal} (hence T_{sys}) as accurately as possible.

At Bure, during an atmospheric calibration, the measured quantities are:

- Phase 1, M_{atm}: the power received from the sky
- Phase 2, M_{load}: the power received from the hot load
- (Phase 3), M_{cold}: the power received from the cold load

T_{rec}, the noise temperature of the receiver, is deduced from the measurements on the hot and cold loads at the beginning of each project and regularly checked. The receiver sideband ratio G^{UL} is also measured at the beginning of each project (see R.Lucas lecture 7). T_{emi}, the effective temperature seen by the antenna, is given by

$$T_{emi} = \frac{(T_{load} + T_{rec}) * M_{atm}}{M_{load}} - T_{rec} \tag{10.7}$$

Moreover, T_{emi} which is measured on the bandwidth of the receiver, can be expressed as the sum of T^L_{emi} and T^U_{emi} (a similar expression exists for T_{sky}):

$$T_{emi} = \frac{T^L_{emi} + T^U_{emi} * G^{UL}}{1 + G^{UL}} \tag{10.8}$$

T_{emi} is directly linked to the sky temperature emissivity (or brightness temperature) T_{sky} by:

$$T_{sky} = \frac{T_{emi} - (1 - \eta_f) \times T_{cab}}{\eta_f} \tag{10.9}$$

were T_{cab} is the physical temperature inside the cabin and η_f, the forward efficiency, which are both known (or measurable) quantities.

Our calibration system provides then a direct measurement of T_{emi} and hence of T_{sky}, which is deduced from quantities accurately measured. Hence, in Eq.10.5 the only unknown parameter remains τ^L, the opacity of the atmosphere at zenith, which is iteratively computed together with T_{atm} the physical atmospheric temperature of the absorbing layers. This calculation is performed by the atmospheric transmission model ATM (see Chapter 8) and the documentation "Amplitude Calibration").

The opacity τ^L (or more generally τ_ν) comes from two terms:

$$\tau_\nu = A_\nu + B_\nu \times w \tag{10.10}$$

A_ν and B_ν are the respective contributions to O_2 and H_2O, the water vapor content w is then adjusted with T_{atm} by the model ATM to match the measured T_{sky}. The ATM model works as long as the hypothesis done on the structure of the atmosphere in plane-parallel layers is justified, as it is usually the case for standard weather conditions.

10.2.1 Low opacity approximation and implication for T_{cal}

When the opacity of the atmosphere is weak ($\tau_\nu < 0.2$) and equal in both image and signal bands, T_{cal} is mostly dependent of T_{atm} and both of them can be considered as independent of τ_ν and hence w.

In the conditions mentioned above, τ_ν can be eliminated from Eq.10.5. The equation becomes:

$$T^L_{cal} = \frac{\eta_F (1 + G^{UL}) \times T_{atm}}{\eta_B \times (1 - \eta_F \times \frac{T_{cab} - T_{atm}}{T_{cab} - T^L_{emi}})} \tag{10.11}$$

(details about the derivation of Eq.10.11 are given in the documentation "Amplitude Calibration" by S.Guilloteau). In Eq.10.11, the unknown is T_{atm}, the physical temperature of the absorbing layers. T_{atm} is mostly dependent on the outside temperature, pressure and site altitude and weakly on τ_ν. For this reason, T_{cal} and T_{sys} remain correct even if w and hence τ_ν are not properly constrained.

Figures 10.1 and 10.2 illustrate this point. Thick lines correspond to the exact equation (Eq.10.5) and dashed lines to the approximation (Eq.10.11). The comparison between Eq.10.11 and 10.5 was done for three common cases 1) at 87 GHz, with $G^{UL} = 10^{-2}$, 2) at 115 GHz, with $G^{LU} = 0.5$ and at 230 GHz, with $G^{UL} = 0.5$. For the 15-m dishes, the forward efficiencies used are $\eta_f = 0.93$ at 3 mm and $\eta_f = 0.89$ at 1.3 mm. Fig.10.1 is done for a source at $elevation = 20^0$ and Fig.10.2 for a source at $elevation = 60^o$.

The following points can be deduced from these figures:

1. As long as $T^L_{sky} = T^U_{sky}$, the equation 10.11 remains valid even at high frequencies > 200 GHz and for $w > 5$ mm.

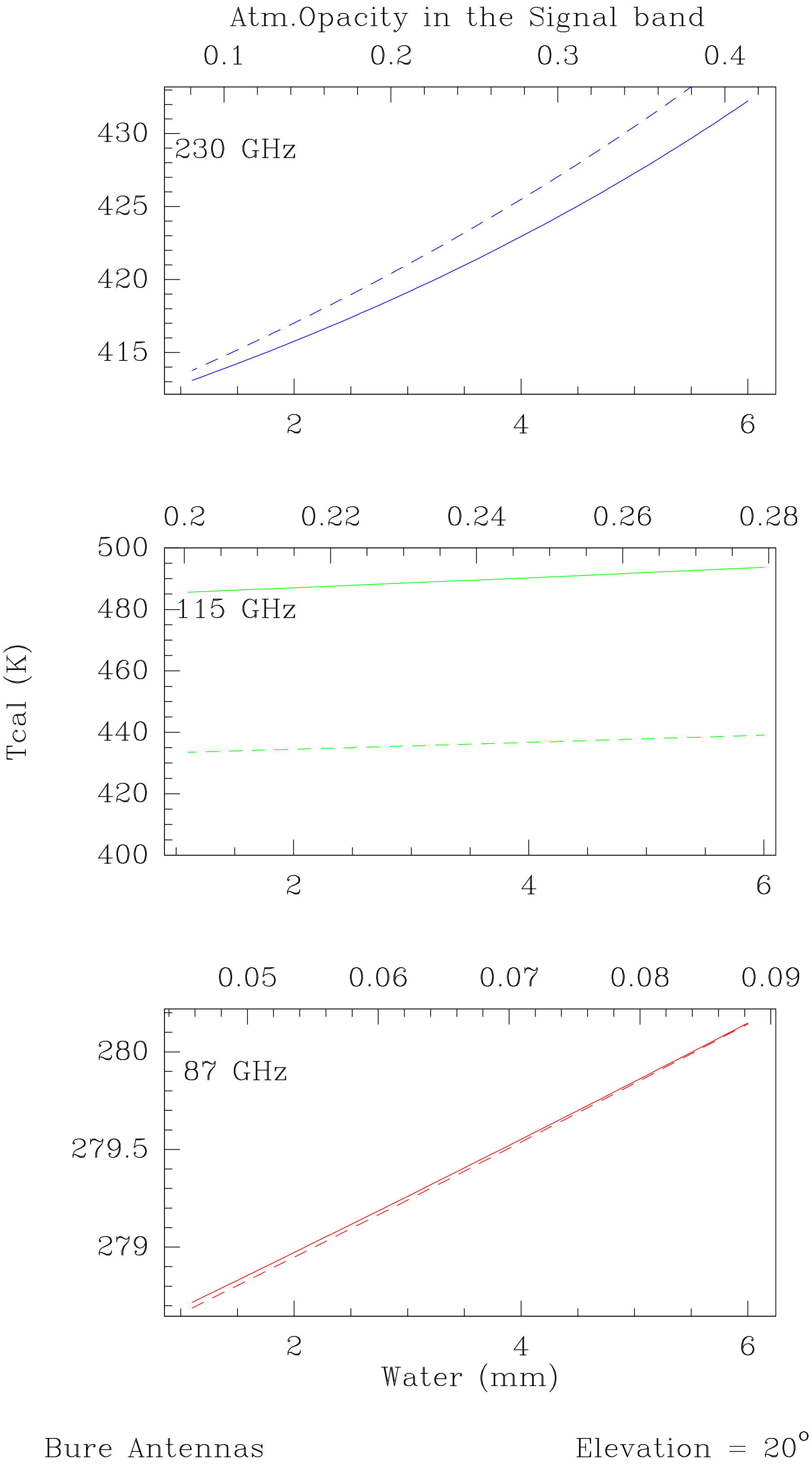


Figure 10.1: Calibration temperature as function of water vapor (or opacity) at 87, 115 and 230 GHz for a source at 20 degrees elevation. Parameters are taken for the Bure interferometer.

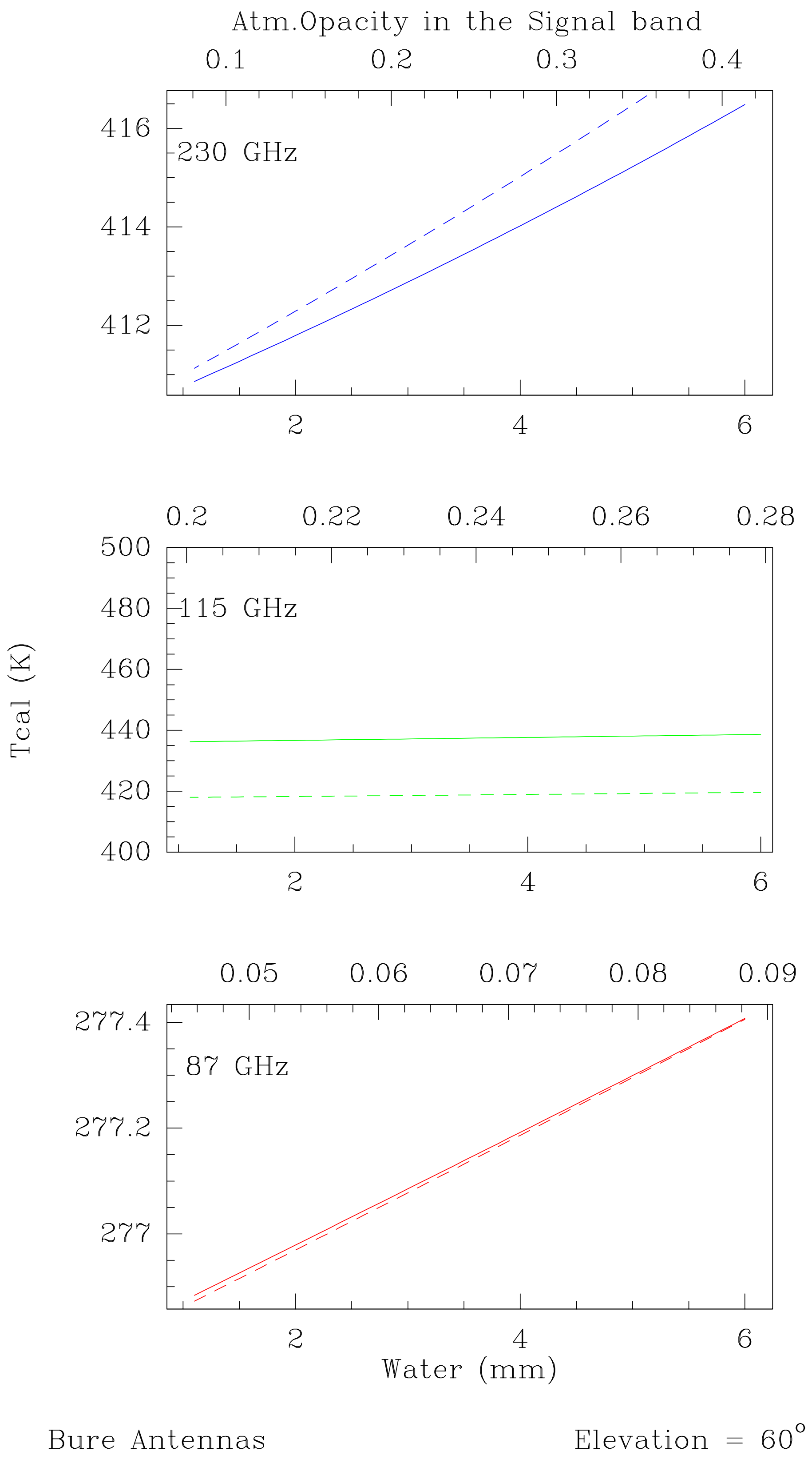


Figure 10.2: Calibration temperature as function of water vapor (or opacity) at 87, 115 and 230 GHz for a source at 60 degrees elevation. Parameters are taken for the Bure interferometer.

Item	T_{cab}	$Trec$	η_f
Typical Error	2 K	10 K	0.01
Induced variation (in %)	0.7	0.3	1.3

Table 10.1: Percentage error on amplitude scale introduced by erroneous input parameters.

2. This comes from the fact the T_{atm} is mostly independent of the atmospheric water vapor content.
3. As soon as $T^L_{sky} \neq T^U_{sky}$, the equation 10.11 is not valid. Note also that the error is about constant with the opacity because T_{atm} is mostly independent of the atmospheric water vapor content. Moreover at 115 GHz, the atmospheric opacity is dominated by the 118 GHz Oxygen line and cannot be below 0.2, the amount of opacity added by the water vapor is small. T_{cal} remains mostly constant with w.

At mm wavelengths, the derivation of the T_{cal} (or T_{sys}) using an atmospheric model is then **quite safe**.

10.2.2 Absolute errors on T_{cal} due to instrumental parameters

The equations above show that T_{cal} is also dependent of the instrumental parameters T_{rec}, η_f and T_{load}. These parameters can also lead to errors on T_{cal}. Derivatives of the appropriate equations are given in the IRAM report "Amplitude Calibration". Applying these equations and taking $T_{atm} = 240$ K, $T_{load} = 290$ K and $T_{emi} = 50$ K, the possible resulting errors are given in the table 10.1.

As a consequence, the most critical parameter of the calibration is the Forward Efficiency η_f. This parameter is a function of frequency, because of optics surface accuracy, but also of the receiver illumination. If η_f is underestimated, T_{sky} is underestimated and you may obtain anomalously low water vapor content, and vice-versa.

The sideband gain ratio G^{UL} is also a critical parameter. G^{UL} is not only a scaling factor (see Eq.10.5), but is also involved in the derivation of the atmospheric model since the contributions from the atmosphere in image and signal bands are considered. This effect is important only if the opacities in both bands are significantly different, as for the J=1-0 line of CO.

Eq.10.5 shows that as soon as the receivers are tuned in single side band ($G^{UL} < 10^{-2}$ or rejection > 20dB), the effect on T^L_{cal} is insignificant. Errors can be significant when the tuning is double-side band with values of G^{UL} around $\sim 0.8 - 0.2$. For example, when the emissivity of the sky is the same in both bands ($T^U_{sky} = T^L_{sky}$), the derivative of Eq.10.8 shows that an error of 0.1 on $G^{UL} = 0.5$ leads to $\frac{\Delta T_{cal}}{T_{cal}} \simeq \frac{dG^{UL}}{1+G^{UL}} \sim 6.5\%$.

However, this problem is only relevant to single-dish observations and cannot happen in interferometry because as soon as three antennas are working, G^{UL} can be accurately measured (see Chapter 7). At Bure the accuracy on G^{UL} is better than about 1 % and the system is stable on scale of several hours.

10.2.3 Relative errors or errors on T^L_{cal}/T^U_{cal}

Following Eq.10.5, the side band ratio will be affected by the following term:

$$T = \frac{e^{(\Delta(\tau^L - \tau^U))}}{\sin(Elevation)} \tag{10.12}$$

where $\Delta(\tau^L - \tau^U)$ is the error on the sideband zenith opacity difference. This difference is maximum at frequencies corresponding to a wing of an atmospheric line, for example when observing around 115 GHz, near the O_2 line at 118 GHz. As example, taking the frequencies of 112 and 115 GHz for a source at 20^o in elevation and a zenith opacity difference $(\tau^L - \tau^U) = 0.150$, an error of 0.030 on this difference (coming from A_ν) will give an error of less than 1% on the gain G^{UL}. Moreover errors on Oxygen lines are very unlikely because the content in Oxygen in the atmosphere is relatively well known and only varying with the altitude of the site.

At the same frequencies, an error of 5 mm (which would be enormous) on the water vapor content will only induce an error of 1% on the gain. Around such low frequency and for small frequency offsets, the water absorption is essentially achromatic. Improper calibration of the water vapor fluctuations will then result in even smaller errors since this is a random effect.

10.2.4 Estimate of the thermal noise

The resulting thermal noise is given by

$$1\sigma = \frac{2kT_{sys}}{\eta A\sqrt{\Delta\nu \times t}}(\mathrm{K}) \tag{10.13}$$

where k is the Boltzmann's constant, A is the geometric collecting area of the telescope, η the global efficiency factor (including decorrelation, quantization, etc...), $\Delta\nu$ the bandwidth in use and t the integration time. The resulting phase noise is inversely proportional to the signal to noise ratio.

10.3 Flux Calibration (visitor's nightmare)

More details are found in the documentation "Flux measurement with the IRAM Plateau de Bure Interferometer"[1] by A.Dutrey & S.Guilloteau.

10.3.1 Introduction

Because of the focus and pointing errors, and possible drifts in receiver gains, amplitude calibration has always been difficult at mm wavelengths. In addition to these basic single-dish effects, the variable amount of decorrelation introduced by phase noise (atmospheric and/or instrumental) make it difficult, if not impossible, for an interferometer to measure absolute flux densities.

All measurements need to be relative to some source of known flux. In practice, planets are used because they are among the few astronomical objects sufficiently strong at millimeter wavelengths for which flux density predictions are possible and sufficiently accurate. They are then used as primary calibrators to bootstrap the flux of the stronger quasars which are point sources. Since the quasars are highly variable, a regular monitoring (each month) is needed. These observations require a very good weather with a small amount of precipitable water vapor (< 4 mm) and a stable atmospheric phase. If not properly taken into account, the quasar variability can produce an error in the flux scale during one configuration which does not result in a simple scale factor in the final image, but introduces artifacts.

10.3.2 Calibration procedure at Bure

Some basic points Because of the physics of quasars, the spectral index may be variable with time as the source intensity. Simultaneous measurements at 2 frequencies are thus needed to estimate it accurately, IRAM instruments (30-m and PdBI) use the frequencies of 86.7 GHz and 228 GHz. At the 30-m, flux density measurements are done during the pointing sessions while they are performed in special sessions at Bure, usually after baseline measurements.

The results of the flux sessions are regularly reduced and published in an internal report (usually each 4 months). These reports are currently available on the web, in the local IRAM page (see[2]).

How we proceed at Bure In practice, it is impossible (and not necessary) to follow all the quasars used as amplitude calibrator at the IRAM interferometer. Monitoring of the RF bandpass calibrators which are strong quasars with flux density > 2 Jy (no more than 4-8 sources) is enough. In the meantime, planets are observed as primary calibrators. These sessions require to calibrate the atmosphere (T_{sys}) on each source and to check regularly the focus.

At the Bure interferometer, the flux density measurements on quasars are done by pointings in interferometric mode. Pointings on planets are actually done in total power mode because they are resolved by interferometry and strong enough. Total power intensity is not affected by the possible decorrelation

[1] http://iram.fr/PDBI/flux/flux.html
[2] http://iram.fr/LI/astro.html

Antenna number	3 mm efficiency (Jy/K)	1.3 mm efficiency (Jy/K)
1	22	37
2	21	27
3	21	36
4	21	29
5	22	34

Table 10.2: Conversion factor from K to Jy for the 15-m antennas of Plateau de Bure

due to atmospheric phase noise. However, it is then necessary to accurately determine the efficiencies of the individual antennas (conversion factor in Jy/K) in interferometric mode ($\mathcal{J}_I$) and in single-dish mode ($\mathcal{J}_S$).

Determining the antenna efficiencies (Jy/K) For each flux session, $\mathcal{J}_S$ is measured on planets by comparison with the models (see GILDAS programs ASTRO or FLUX).

For a given antenna, the interferometric efficiency $\mathcal{J}_I$ is always $\geq \mathcal{J}_S$. Pointings measurements in interferometric mode are not limited by the atmospheric decorrelation because the timescale of the atmospheric decorrelation is usually significantly larger than the time duration of the basic pointing integration time (< a few sec). On the contrary, all instrumental phase noise on very short timescale can introduce a significant decorrelation and degrades $\mathcal{J}_I$. This is what may happen from time to time at a peculiar frequency due to a bad optimization of the receiver tuning.

For example, in the initial 1.3 mm observations, strong decorrelation was introduced by the harmonic mixer of the local oscillator system which degraded $\mathcal{J}_I$ by a factor of $2-4$ depending of the antennas. This problem has been solved recently. Now at 3 mm, it is reasonable to neglect the instrumental noises and take $\mathcal{J}_I = \mathcal{J}_S$. At 1.3 mm, the new harmonic mixers have been installed only recently and statistics on the site are rare but laboratory measurements show that the loss in efficiency should be small. The Table 10.2 gives the antenna efficiencies $\mathcal{J}_S$, as measured in flux sessions or by holography. These values are the current efficiencies (since November 1997); older values are given in flux reports. They assume that the focus is optimum and do not include any instrumental phase noise. $\mathcal{J}_I$ agrees usually within 10 % at 3 mm and 15 % at 1.3 mm with $\mathcal{J}_S$, note that $\mathcal{J}_I$ must be $\geq \mathcal{J}_S$.

Being able to cancel out most of the instrumental phase noise even at 1.3 mm makes the IRAM interferometer a very reliable instrument. It is reasonable to think that, in the near future, the flux calibration will be systematically performed at Bure at the beginning of each project by reference to the antenna efficiencies. This is indeed already the case: after pointing and focusing, we systematically measure the flux of calibrators when starting a new project (data labeled `FLUX` in files). Up to now, for typical weather conditions, most (more than 90 %) of the flux measured at 3 mm are correct within 10 % and more than 60 % at 1.3 mm are within 15 %.

CRL 618 and MWC 349 as secondary flux calibrators Finally, for each project, a complementary flux check is systematically done using the continuum sources CRL 618 or MWC 349 (pointing + cross-correlations). However these sources must be used with some caution. CRL 618 is partially resolved in A and B configurations at 3 mm and in A,B,C at 1.3 mm. Moreover it has strong spectral lines which may dominate the average continuum flux; this must be checked before using it for flux estimates. MWC 349 is unresolved and remains a reliable reference in all antenna configurations. The only strong lines for MWC 349 are the Hydrogen recombination lines. The adopted flux densities are:

For CRL618 (see flux reports 13 and 15):

- CRL618 F(87 GHz) = 1.55 Jy (+/ − 0.15)
  At 87 GHz, the flux density of CRL 618 (free-free emission from the `HII` region) has increased since 1990 (where it was ∼ 1.1 Jy instead of 1.55 Jy).
- CRL618 F(231.9 GHz) = 2.0 Jy (+/ − 0.3),
  from [Martin-Pintado et al 1988]. Since the flux density has increased at 87 GHz, this value needs

to be observationally confirmed. Beware of the line contamination which can be high in CRL618.

For MWC349:

- Spectrum of MWC349 $F(\nu) = 1.69(\nu/227\ \mathrm{GHz})^{0.6}$
- MWC349 F(87 GHz) = 0.95 Jy
- MWC349 F(227 GHz) = 1.69 Jy

These values agree within 1 σ with the measurement performed at 87 GHz [Altenoff et al 1994], (0.87 $\pm$0.09 Jy).

10.3.3 Determining the absolute flux scale on a project

The method Fig.10.3 is a printout of the "standard calibration procedure" used in CLICThis procedure uses the CLIC command `SOLVE FLUX` which works on cross-correlation only as follows:

1. The flux of the reference source is fixed to $F(\mathrm{Ref})$
2. $F(\mathrm{Ref})$ is used to measure the antenna efficiency by dividing by antenna temperature of the reference ($T_A^*(\mathrm{Ref})$): $\mathcal{J}_I = T_A^*(\mathrm{Ref})/F(\mathrm{Ref})$
3. $\mathcal{J}_I$ is used to compute the flux of all other sources in the index: $F(source) = \mathcal{J}_I \times T_A^*(source)$

The flux density of the amplitude calibrators will be used in the final step of the amplitude calibration to fix the flux of the source of astronomical interest.

The practice In the automatic procedure, the reference sources are the calibrators where **Fixed flux** is set to **YES** and the reference values are in the variable **Input Flux**. **Flux in file** corresponds to the value stored with the data (by using the observational command `FLUX`, see lecture 6 by R.Neri for details). The calculation is performed by clicking on `SOLVE` and the results are displayed inside the variable **Solved Flux**.

If you want to iterate using one of these values as reference, you need to write it in the variable **Input Flux** and set **Fixed flux** to **YES**. Like in the CLIC command `SOLVE FLUX`, the individual antenna efficiencies ($\mathcal{J}_I$) are computed; these values are only averaged values on the time interval using all sources. They are then affected by many small biases like pointing or focus errors and atmospheric decorrelation. They are then usually worse than the canonical values given in table above (for biases, see end of this section).

When you are satisfied by the flux calibration, you need to click on the following sequence of buttons: 1) `Get Results` in order to update the internal variables of the CLIC procedure, 2) `Store` to save the flux values inside the header file (`hpb` file) and 3) `Plot` to display the result of your calibration. The plot shows the inverse of the antenna efficiencies ($1/\mathcal{J}_I$) versus time for all selected sources. If the flux calibration is correct, all sources must have the same value *e.g.* $1/\mathcal{J}_I$. This plot is systematically done in mode `amplitude scaled` (written on the top left corner). In this mode, the antenna temperature of each source $T_A^*(source)$ in K is divided by its assumed (variable **Input flux**) flux density $F(source)$ in Jy (the value you have just stored), the result is then $T_A(source)^*/F(source) = 1/\mathcal{J}_I(source)$ which must be the same for all sources and equal to $1/\mathcal{J}_I$. If it is not the case, for example if one source appears systematically lower or higher than the others, this means that its flux is wrong and you need to iterate.

Note that the scan range, applied on all calibrators, which is by default the scan range of the "standard calibration procedure" can be changed. This option is useful when there is some shadowing on one calibrator because the shadowing can strongly affect the result of a `SOLVE FLUX`. If you change the scan range, do not forget to click on `UPDATE`.

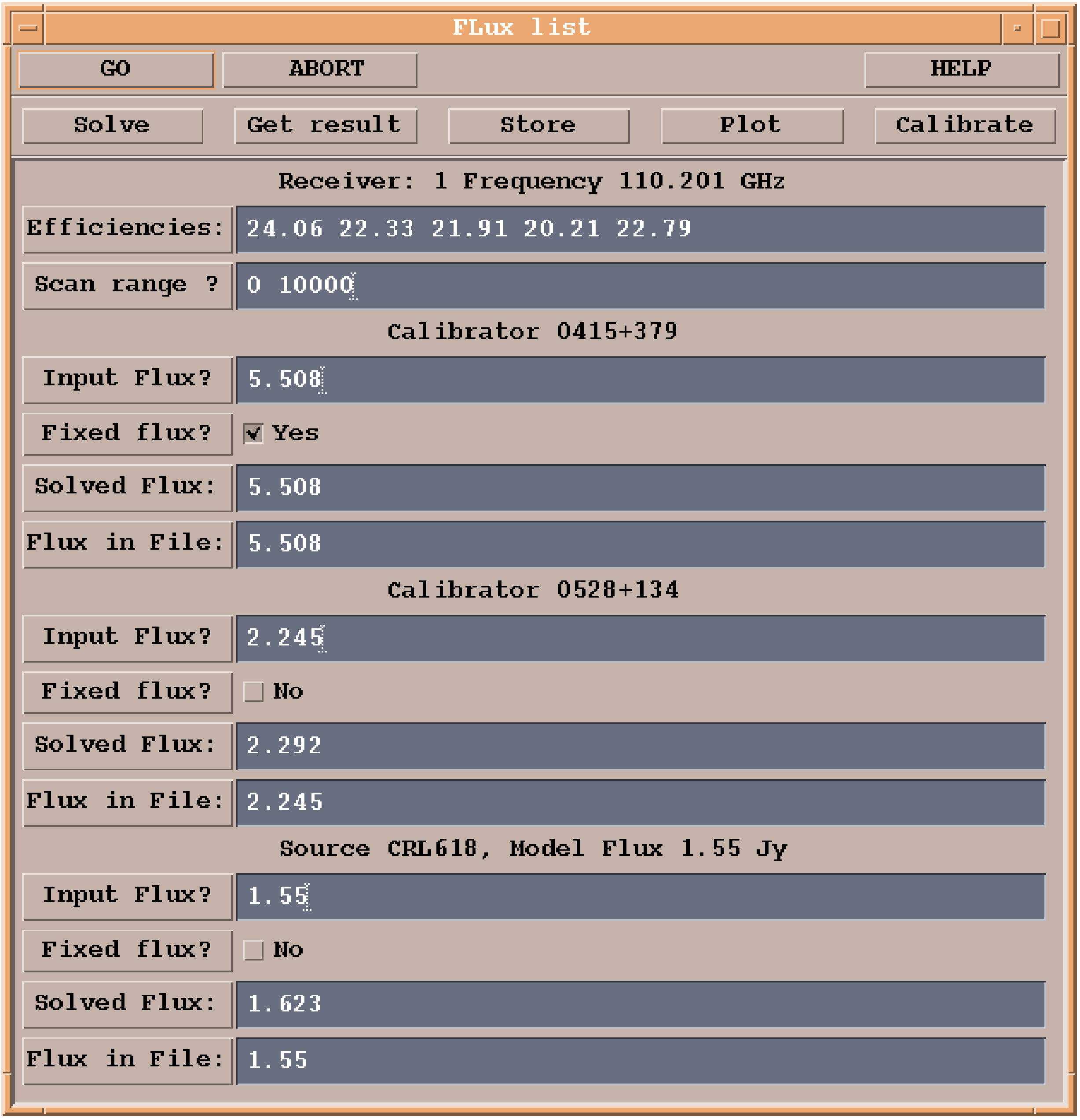


Figure 10.3: User interface of the "standard calibration procedure" of CLIC corresponding to the flux calibration.

10.3.4 Possible biases and remedies

Flux densities are more important than efficiencies In the final amplitude calibration performed on the source (see next section), the flux of the source is determined by reference to the flux of the amplitude calibrator which is usually also the phase calibrator. This means that the averaged efficiencies $\mathcal{J}_I$ computed by `SOLVE FLUX` and the automatic procedure are not directly used and in many case variations of $\mathcal{J}_I$ does not affect the accuracy of the final amplitude calibration because they are corrected. It is then fundamental to have a good estimate of the flux of the amplitude calibrator but not necessarily to know precisely the averaged $\mathcal{J}_I$.

Possible biases Using the automatic procedure, the possible biases are the following:

1. There is some shadowing on the reference source, the estimate of the $\mathcal{J}_I$ can be wrong. Use another reference.

2. One or several antennas are off focus: $\mathcal{J}_I$ is larger than $\mathcal{J}_S$ but flux densities can still be correct if there is no significant drift during the time interval used to measure the fluxes. If the data are affected by a significant focus drift, it also affects the accuracy of the flux measurements. Depending of the observation time of the reference and of the sources, the estimated flux densities can be either too low (reference taken at the beginning when the focus is correct, sources at the end when the focus is off) or too high (opposite situation). In both cases, it is necessary to check the focus (data called **FOCUS** or have a look to the `show.ps` file). In the first case, the measured fluxes are correct. In the second case, the flux calibration must be done on a smaller interval of time where the focus drifts remain negligible.

3. The pointing on the reference is bad, $\mathcal{J}_I$ is overestimated implying that the flux of all other sources (with good pointing) is also overestimated. Check the pointing on the possible reference sources (data called **POINT** or have a look to the `show.ps` file) and select a better reference.

4. There is a strong atmospheric decorrelation. Flux measurements are performed on cross-correlations of about 4 minutes and the atmospheric phase fluctuations are high (check them on an individual cross-correlation taken on a strong quasar *e.g.* the RF calibrator), there are two possibilities: i) the atmospheric correction works well (clear sky), this is usually the case and it is necessary to apply it to measure the fluxes or 2) the atmospheric correction does not work (cloudy sky), the data can be used at 3 mm and the flux scale can be computed by selecting the best scans on a small interval of time but at 1.3 mm data are useless.

5. The interferometric efficiencies $\mathcal{J}_I$ are really very different to $\mathcal{J}_S$ because there is a wonderful mixing of the points mentioned above... Ask to an expert (your local contact first, myself later if needed...).

Note that the biases 3) and 4) do not affect flux estimates when they are performed on pointing data (like in the case of the flux sessions).

10.3.5 The program FLUX

This program is not used by the external users of the PdBI but IRAM astronomers to provide reliable flux density of quasars to visitors (see flux reports). A description of the program is given in "Flux measurement with the IRAM Plateau de Bure Interferometer"[3] by A.Dutrey & S.Guilloteau

10.4 Interferometric Calibration of the Amplitude

For antenna i, the antenna-based amplitude correction is given by (Eq.10.3 and 10.4).

$$a_i^K(t) = T_{cal\,i}^K(t) G_i^K(\nu, t) \mathcal{B}_i(t) \tag{10.14}$$

[3] http://iram.fr/PDBI/flux/flux.html

where K = U or L. The decorrelation factor f (see R.Lucas lecture 7) is not taken into account here because it is fundamentally a baseline-based parameter.

In a baseline-based decomposition, the complex gain of baseline ij, $\mathcal{G}_{ij}$ is given by:

$$\mathcal{G}_{ij}^K(t) = f \times a_i(t)a_j(t)e^{i(\phi_i(t)-\phi_j(t))} \tag{10.15}$$

and the amplitude of the baseline ij is $\mathcal{A}_{ij}$

$$\mathcal{A}_{ij}^K(t) = f\sqrt{T_{cal\,i}^K T_{cal\,j}^K(t)G_i^K(\nu,t)G_j^K(\nu,t)\mathcal{B}_i(t)\mathcal{B}_j(t)} \tag{10.16}$$

We will discuss first the term $\mathcal{B}_i$ and estimate then the decorrelation factor f, before giving a global scheme of the amplitude calibration.

10.4.1 Correction for the antenna gain $\mathcal{B}_i(t)$

The antenna gain $\mathcal{B}_i(t)$ corresponds to losses due to the antenna, mainly focus (F_i) and pointing (P_i) errors coming from thermal variations of the antenna structure and surface.

$$\mathcal{B}_i(t) = P_i(t) \times F_i(t) \tag{10.17}$$

At Bure, we now check and correct automatically the pointing and the focus each hour. This correction is then done mainly in real time. This has improved a lot the quality of the data at 1.3 mm. However, it is necessary in some cases to add a break in the amplitude (but not in the phase) fitting in order to take into account a focus error or a loss of amplitude due to pointing errors.

Note that, an error on the focus of 0.1 mm at 1.3 mm will introduce a phase error of $\frac{4\pi}{\lambda} \times \frac{180}{\pi} \sim 24^o$ and a loss in amplitude of $\sim 3-5\%$.

10.4.2 Estimate of the atmospheric decorrelation factor f

Details about the origin of f are given in M.Bremer lecture 9. I will discuss here the practical implementation of the atmospheric phase correction done in real-time and in CLIC. More details are given in the IRAM report "Practical implementation of the atmospheric phase correction for the PdBI" by R.Lucas.

The atmospheric phase fluctuations are due to different time varying water vapor content in the line-of-sight of each antenna through the atmosphere. Between antenna i and j, this introduces a decorrelation factor $f \sim e^{-(\phi_i-\phi_j)^2/2}$ on the visibility V_{ij}. This term, non-linear, cannot be factorized by antenna. Moreover due to the physical properties of the atmosphere, there are several timescales. One can correct partially some, but not all, of them.

At Bure the basic integration time is 1 second and the scan duration is usually 60 seconds. The radiometric correction works then on timescales of a few seconds to one minute. It corrects only the amplitude: the phase is never changed because phase jumps between individual scans are dominated by instrumental limitations (mainly the receiver stability on a few minutes + ground pickup variations). The implications on the image quality are developed in the lecture by S.Guilloteau 16. Longer atmospheric timescales of about $2-8$ hours are removed by the spline functions fitted inside the phase and the amplitude.

Intermediate timescales fluctuations from about one minute (the scan duration) to 1 hour are not removed. The resulting rms phase are measured by the fit of the splines in the phase. These timescales are not suppressed by the radiometric correction, and they contribute to the decorrelation factor f (see Eq.10.16). as the main component.

The method The differences in water vapor content are measurable by monitoring the variations of the sky emissivity T_{sky}. A monitoring of the total power in front of each antenna will then lead to a monitoring of the phase fluctuations. At Bure, we monitor the total power P with the 1.3 mm receivers (note that P is also called M_{atm} in the first part of this lecture). The variation of T_{sky}, ΔT_{sky} (equal to ΔT_{emi}) is linked to the total power by

$$\frac{\Delta P}{P} = \frac{(\Delta T_{emi} + \Delta T_{loss})}{T_{sys}} \tag{10.18}$$

The monitoring of the atmospheric phase fluctuation works only when ΔT_{loss} due to the instrumentation is negligible on the time scales at which the phase correction is calculated and applied (typically a few seconds to one minute). Slow drifts on scale of hours have no effects.

With standard atmospheric conditions and following [Thompson et al 1986] (their Eq.13.20), the variation of the path length through the atmosphere at zenith is approximated by:

$$\Delta L = 6.3\delta w \tag{10.19}$$

were δw is the variation of water vapor content. ΔL is related to the phase fluctuation ψ_i above the antenna i by

$$\psi_i(t) = \frac{2\pi}{\lambda}\Delta L(t) \tag{10.20}$$

For example, under standard conditions (see fig.10.1 or 10.2), a variation $\delta w = 0.1$ mm corresponds to $\Delta L \simeq 630$ μm, $\Delta T_{sky} \simeq 1.5$ K and $\psi_i \simeq 250^o$ at 1.3 mm. This value is enormous and would not allow to produce images of good quality.

To reduce the phase fluctuation to a reasonable value having a negligible impact on the image quality *e.g.* $\psi_i \sim 25^o$, one needs to get $\Delta T_{loss} + \Delta T_{sky} \sim 0.15$ K corresponding to a global path length variation of $\sim 60\mu m$. For a typical $T_{sys} \sim 150$ K (DSB in the antenna plane, not SSB outside the atmosphere as for astronomical use), the instrumental stability required $(\Delta T_{loss}/T_{sys})$ must then be of order of $\sim 5 \cdot 10^{-4}$.

At Bure, on time scales of a few minutes, ΔT_{loss} is dominated by the stability of the receivers which must be carefully tuned to get the best stability. The 1.3 mm receivers are systematically tuned to get a stability of a few 10^{-4}; the stability is checked by doing autocorrelations of 60 seconds on the hot load. Achieving the required stability may prove impossible at some frequencies.

Practical implementation Ideally one would like to use T_{emi} measured each second on each antenna to compute $\psi_i(t)$ and correct the measured baseline phases. Practically, it is not so simple because $\psi_i(t)$ can do many turns and instrumental effects affect the measured T_{emi}.

Instead we use a differential procedure: once the antenna tracks a given source, one calibrates the atmosphere to calculate $T_{sys}(t_0)$, $\Delta L(t_0)$ and $\Delta L/dT_{sky}(t_0)$. Phase corrections are then referenced to t_0.

$$\Delta\psi_i = \frac{2\pi}{\lambda}\frac{d\Delta L}{dT_{emi}}\frac{T_{sys}(t_0)}{P(t_0)}(P(t) - P(\mathrm{Ref})) \tag{10.21}$$

where $P(\mathrm{Ref})$ is chosen in order to minimize as much as possible all the slow effects contributing to ΔT_{loss}. A long term atmospheric effect can also be included in $P(\mathrm{Ref})$ because these effects are not removed by the radiometric phase correction but by the traditional phase referencing on a nearby calibrator. The main steps are the following:

1. The total power P is continuously monitored on calibrators and on sources (each one second).

2. Using the standard calibration method (see first part of the lecture) P and T_{emi} (measured each second) are used to compute T_{sky} and w.

3. The atmospheric model has also been upgraded to compute the path length ΔL and its derivative $d\Delta L/dT_{emi}$. ΔL is computed by integrating the refractive index of the wet air along the line of sight across the atmosphere.

4. Inside the 60 seconds scan, the new phase (Eq.10.21) is computed and the correction applied to the amplitude.

Quasi-real Time Calibration For the quasi-real time correction,

- the default value for $P(Ref)$ is the measured atmospheric emission at the time of the last calibration, i.e. $P(\mathrm{Ref}) = P(t_0)$.

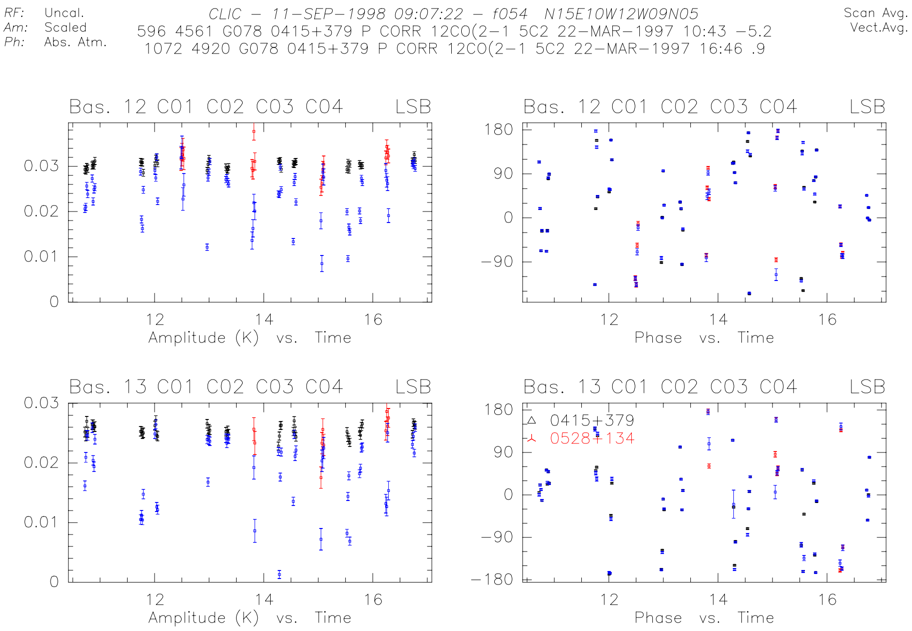


Figure 10.4: The amplitude and phase versus time on baselines B12 and B13 with (black) and without (red and blue) the radiometric phase correction. The phase remains unchanged but the amplitude is significantly improved.

Calibrating using CLIC The CLIC command "**MONITOR** *delta – time*" allows to re-compute all the parameters. This command is useful when you want to select a better value for $P(\mathrm{Ref})$.

- This command is used to prepare the atmospheric radiometric phase correction. It processes the calibration scans to compute the correction factors (i.e. the change of path length for a given change in emission temperature of the atmosphere at the atmospheric monitor frequency (normally 1.3 mm)).
- The scans in the current index are grouped in intervals of maximum duration *delta – time* (in seconds); source changes will also be used to separate intervals. In each interval a straight line is fitted in the variation of atmospheric emission as a function of time; this line will be the reference value for the atmospheric correction, i.e. the correction at time t is proportional to the difference between the atmospheric emission at time t and the reference at time t. This scheme is used to avoid contaminating the correction with total power drifts of non-atmospheric origin (changes in receiver noise and gain, and changes in ground noise).
- `MONITOR 0` will use for each scan the average of the atmospheric emission as the reference value (i.e. $P(\mathrm{Ref}) =< P(t) >_{scan}$). This will cause the correction to average to zero in one scan: the average phase is not changed, only the coherence is restored leading to an improved amplitude.

The automatic calibration procedure uses the command `MONITOR 0`.

10.4.3 Fitting Splines: the last step

In the real-time processing, only the receiver gain and bandpass, the atmospheric transmission and the radiometric correction have been calibrated.

Fitting of the temporal variations of the global antenna gain (the so-called *amplitude calibration*) is performed in CLIC by fitting splines functions with time steps of 3-6 hours (`SOLVE AMPLITUDE [/WEIGHT] [/POL` *degree*`] [/BREAK` *time*`]`) and can be done either in baseline-based or in antenna-based mode. Note that in the latter case, the averaged amplitude closures are computed, as well as their standard deviations. The amplitude closures should be close to 100%. Strong deviations of amplitude closures from 100% are an indication of amplitude loss on long baselines, due to phase decorrelation during the time averaging.

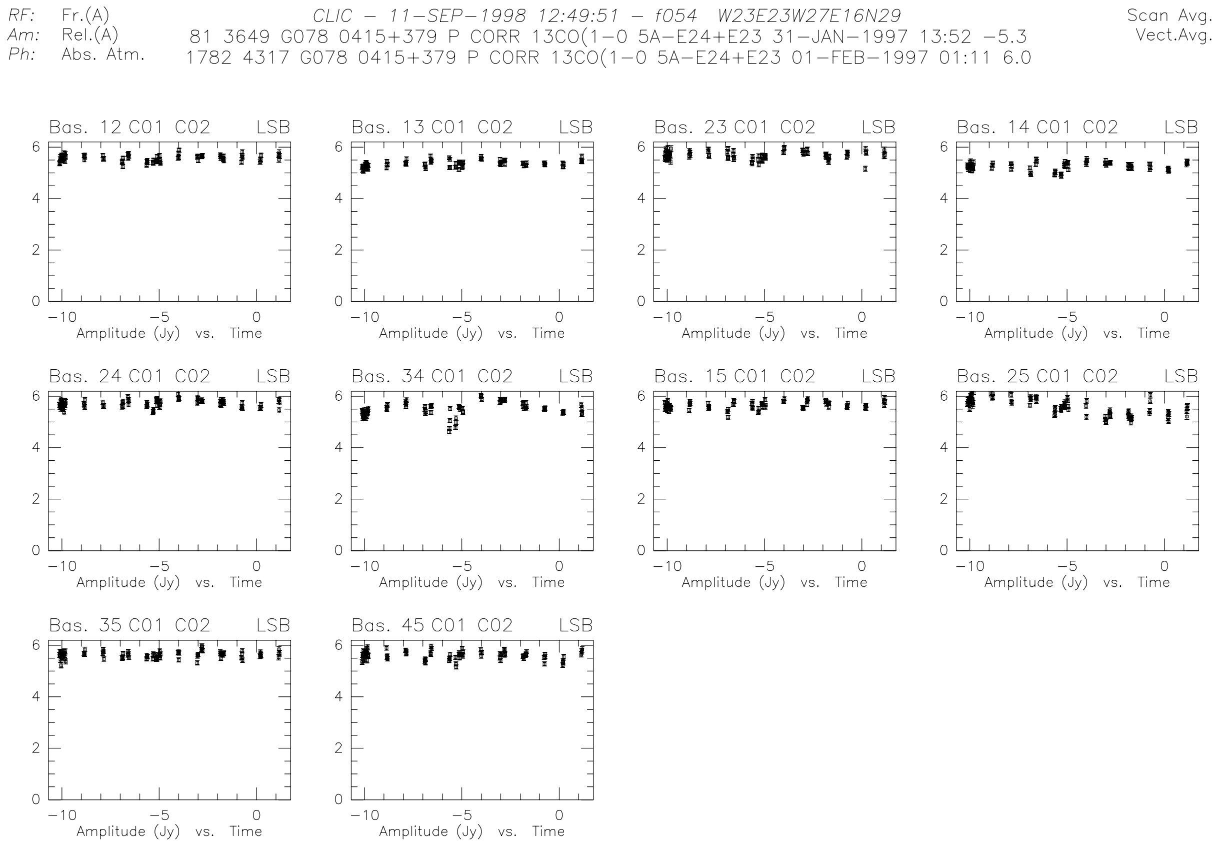


Figure 10.5: Calibrated amplitude (Jy) versus time

The fit then shows systematic errors; if this occurs, baseline based calibration of the amplitudes might be preferred.

The amplitude calibration involves interpolating the time variations of the antenna gains measured with the amplitude calibrator, assuming the its flux is known. The fitted splines must be as smoothed as possible in order to minimize the errors introduced on the source which is observed in between calibrators.

10.4.4 A few final checks

Once the amplitude calibration curve is stored, one can perform some simple checks on the calibrated data of the calibrator. These checks must be done in Jy (mode "`AMPLITUDE ABSOLUTE RELATIVE`" to the flux density of the calibrator).

Amplitude versus time On each baseline, the amplitude curves should be flat and equal to assume the flux density of the calibrator.

Amplitude versus IF frequency On each baseline, the amplitude curves should be flat, but they are not necessarily equal to flux of the calibrator because the decorrelation factor f is not taken into account here. To retrieve the flux density of the calibrator, they must be multiplied by the corresponding $e^{-(\Delta\phi)/^2}$, where $\Delta\phi$ is baseline rms phase noise determined during the phase calibration.

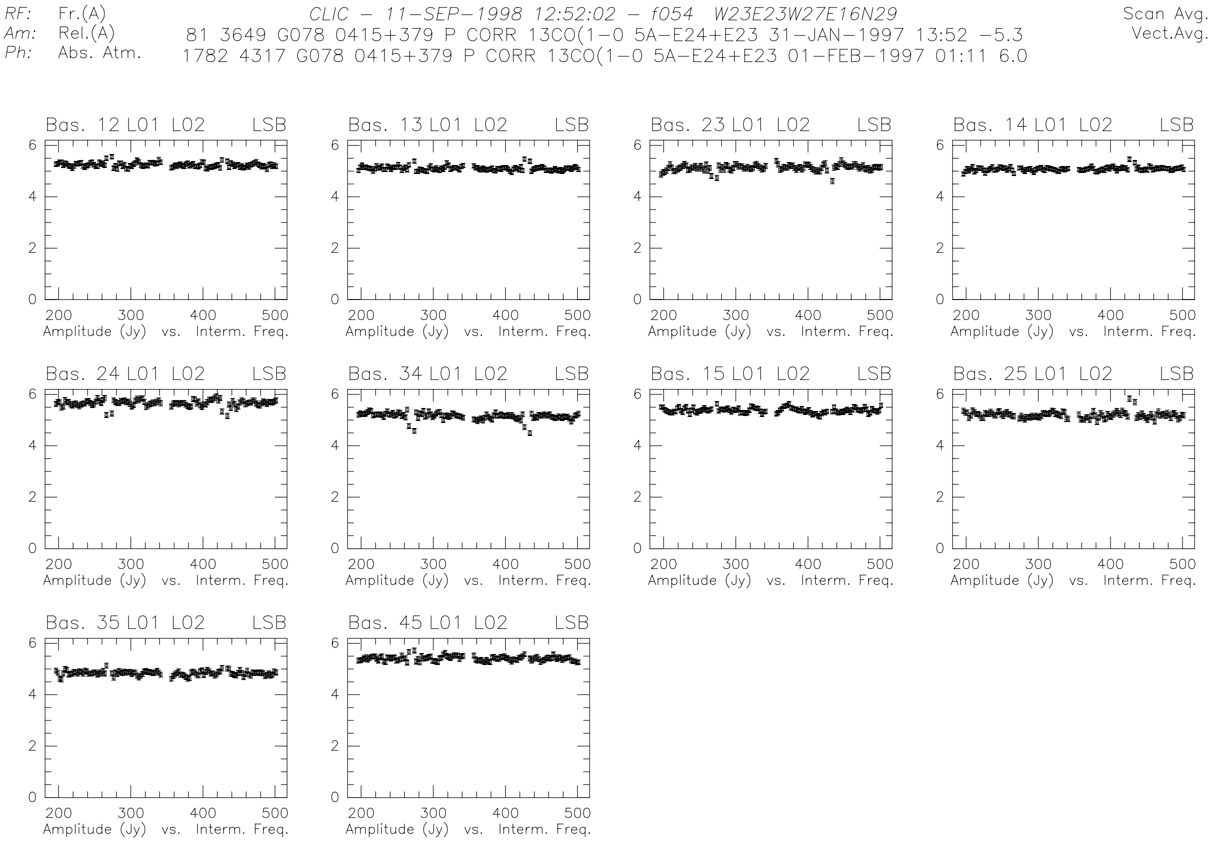


Figure 10.6: Calibrated amplitude versus IF frequency. The decorrelation factor is not calibrated out, and varies from baseline to baseline.

Chapter 11

Calibration in Practice

Anne Dutrey & Roberto Neri

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

11.1 Introduction

11.1.1 Contents of the account

At the first login on a new project account, you should find five directories:

1. `reports:` It contains all the pre-calibration performed by the Astronomer on Duty (AoD) on the site. This subdirectory is in read access only. For a given observing date, you should find the following compressed files: `28-feb-2001-x007.hpb.gz`, the header file containing the calibration curves of the data; `28-feb-2001-x007.ps.gz`, the PostScript file containing the result of this calibration. This file should be carefully read before starting a new calibration. `show-28-feb-2001-x007.ps.gz` is the PostScript file corresponding to a first look of the data. It includes meteo conditions (wind, mm of water, system temperature) but also pointing and focus errors. In many cases, this file can help to interpret dubious data. This file should also be read carefully before starting calibration. The file `x007.note` are notes written by the AoD on the site. This file is updated at each period of observations. It gives important information about the data quality and possible problems.

2. `headers:` It initially contains a copy of reports. Files should be uncompressed and data calibration should be performed in this subdirectory. This is the default directory of CLIC. However before activating CLIC, it is better to move to this subdirectory using the shell command: `‘‘cd /headers’’`.

3. `maps:` This subdirectory, empty at the beginning, should contain at the end all the *uv* tables and the maps produced. Using the CLICprocedure `‘‘Write A UV Table’’`, the *uv* tables are created by default in this directory which is also the default directory for GRAPHIC and MAPPING.

4. `tmp/DATA`: This subdirectory contains the raw data files (IPB) and their corresponding log files. In the example presented above, it would contain: `H228X007.IPB, H228X007.OBS, H228X007.RDI`. Access to this directory and to the raw data files is automatically handled by CLIC.

5. `DATA:` Most of the time this directory is empty. However, it can contain some IPB files and their corresponding log files. Like in the case of `tmp/DATA`, its use is transparent. The reducer does not need to know the exact location of the raw data files on her/his account. CLIC handles it for him/her.

Note that only the directories `reports, headers` and `maps` are saved on CDroms by the automatic procedure `savproj`.

11.1.2 Before starting the data reduction

The compressed PostScript files in `headers` containing the calibration curves performed by the AoD must be *gunzip*ped and printed. Then, they must be carefully studied in order to determine the general quality of the data and find possible problems which are usually mentioned in the `project.note` file (*e.g.* `x007.note`) written by the AoD during the pre-data reduction process.

11.1.3 Activating the CLIC environment

CLIC is available by typing in the directory `headers`

```
$ CLIC
```

Under CLIC, there are several procedures available:

1. `File`: To open a new header (hpb) or data (IPB) file
2. `Raw Data File`: To open a raw data file (IPB)
3. `Raw Data File Directories`: To define the location of raw data file directories
4. `First Look`: Instrumental and meteorological conditions during the observations (see §2)
5. `Standard Calibration (2 receivers)`: To calibrate raw data (see §3)
6. `Self-Cal on a point source`: To self calibrate the phase
7. `Simple Check`: To do simple checks on the data
8. `Holography Reduction`: To reduce holographic data
9. `Write A UV Table`: To produce the visibility table

Useful procedures to calibrate data are 1) `‘‘First Look’’`, 2) `‘‘Standard Calibration (2 receivers) ’’` and 3) `‘‘Write A UV Table’’`.

11.2 The “First Look” procedure

The procedure `‘‘First Look’’` provides information about the weather conditions and a few instrumental parameters at the time of the observations. This information is very important as it helps you to make a first data quality assessment even before you may start with the interactive data calibration procedure. The panel of the procedure is shown on Fig.11.1.

Monitoring information is provided on:

- `Meteorology:` the average and maximum wind velocity, the ambient pressure and temperature. Gusty conditions and observations with wind velocities above 10 m/s may show up with high pointing corrections. Take care to tag visibilities which may be affected by such difficult observing conditions.
- `Pointing and Focus:` the applied corrections are shown for all the antennas in the array. Only differential variations in the corrections play a role, not the absolute amount. Sudden pointing corrections by more than 10″ can considerably spoil the visibilities, especially at the highest observing frequencies. A similar consideration applies for antenna focus corrections, although visibilities are automatically corrected for phase offsets which are generated by focus corrections.

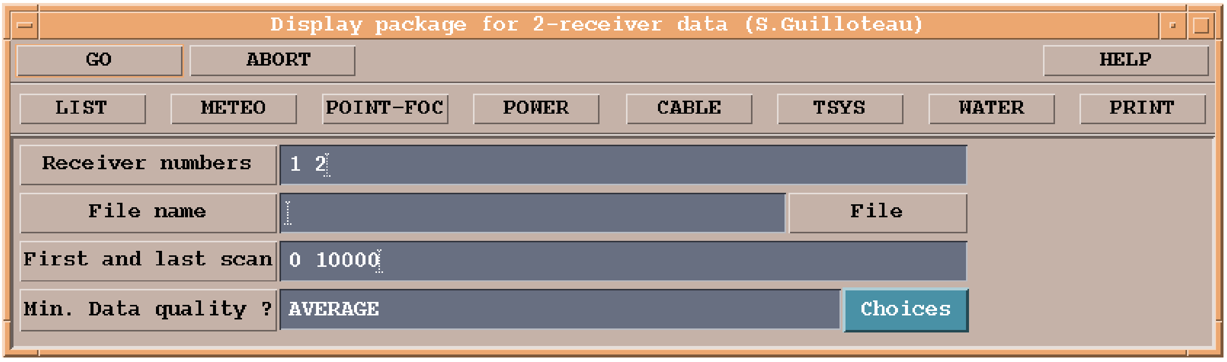


Figure 11.1: “First Look” panel

- `Total Power:` a trace of the total incident atmospheric power recorded by the continuum detectors (one for each receiver, each one second), helps in further evaluating the data quality. As a rule, strong and rapid variations in the total power trace the presence of clouds in the line of sight while a sudden up- or down-stepping on one antenna and on one receiver is a sign of a renewed receiver tuning. The total power increases in general with the air mass.

- `Cable Phase:` variations in the electrical length of the cables show up in phase variations (referenced to the LO2 phases). These are measured by a phasemeter. Appropriate corrections for the phases of the LO1 rotators are computed taking $n$times the corrections measured by the phasemeters of the 3 mm receivers and $3\,n$ times the corrections measured by the phasemeters of the 1 mm receivers, where n is the harmonic number of the tuned frequency. Strong and rapid variations while a source is tracked may indicate a fault in a cable (these data should perhaps be flagged), whereas a sudden but steady change is mostly related to a shift in the antenna pointing.

- `System Temperature:` dependent on the observing conditions and on the frequency. As for the total power detectors, strong and rapid variations in the system temperature trace atmospheric instability, whereas a sudden but steady change on one receiver is a sign of a renewed receiver tuning.

- `Water vapor:` the content of precipitable water vapor in the atmosphere is a critical parameter on which the quality of most of the high frequency observations depend. Should the measured water vapor content not be consistent (or roughly) on all the high frequency receivers, please check the receiver gain.

Finally, the “First look” procedure produces a short list of all the scans recorded at the time of the execution of a project. Such a listing allows you to trace back the sequence of operations during an observing run. Note that the range of scan validity for the calibration procedure sets up with the last `GAIN` scan in the short list.

Looking at the results of the procedure (called in the example above `show-28-feb-2001-x007.ps.gz`) should be done simultaneously with the reading of the project.note file (here `x007.note`).

11.3 The “Standard Calibration (2-receivers)” procedure

We describe here, step by step the inputs and actions (or outputs) of the procedure `‘‘Standard Calibration (2 receivers) ’’`. The associated panel is given in Fig.11.2.

11.3.1 Inputs

On the panel, the reducer should select with the mouse the `‘‘File name’’` of the header file. Other parameters are automatically selected by the procedure if the flag `‘‘Use previous settings’’` is set to `YES` which is the default. Never change it if your data are correct (no editing). The parameters `‘‘First`

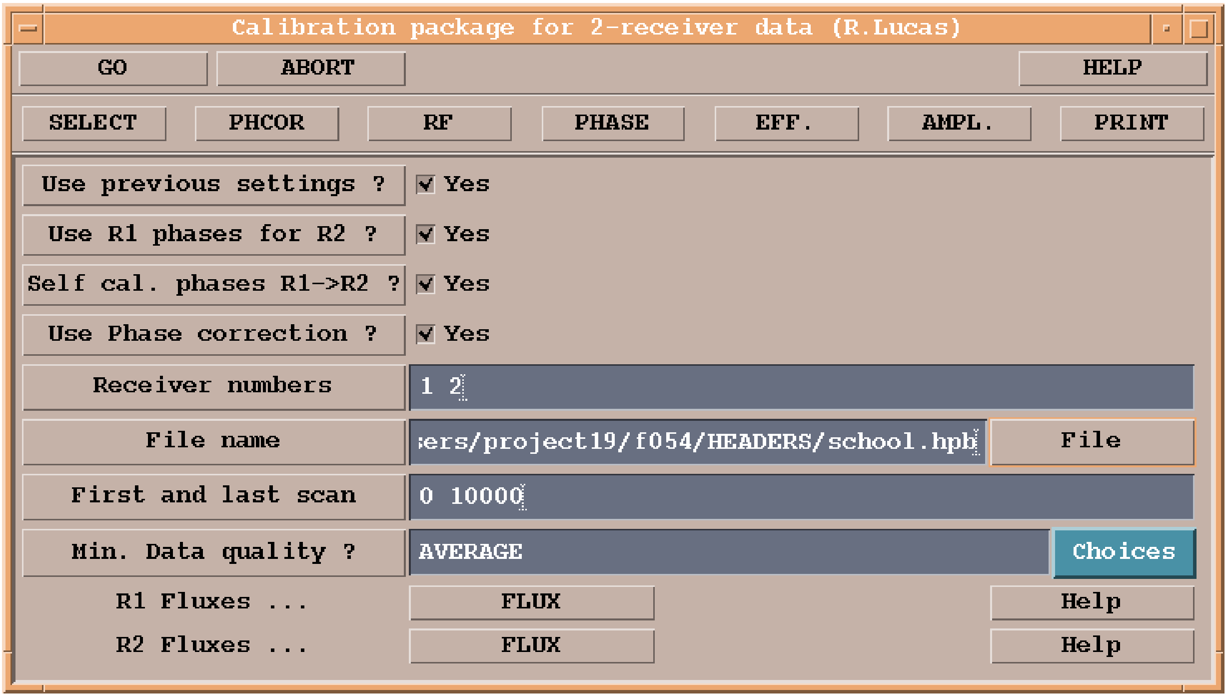


Figure 11.2: “Standard Calibration (2 receivers)” panel

and last scan’’ are selected automatically when ‘‘Use previous settings’’ is set to YES. The default value for the flag Min.Data quality is AVERAGE. The flag ‘‘Use Phase correction’’ must also be set to the default: YES. The current procedure uses by default the phase calibration of the receiver 1 to calibrate the instrumental phase of the receiver 2 because the experience has shown that it is the more efficient way to proceed (see R.Lucas lecture on phase calibration). Therefore, the calibration of the receiver 2 (1.3 mm) cannot be dissociated from the calibration of the receiver 1 (3 mm) and the flag ‘‘Receivers numbers’’ must be set to 1 2, except when observations were done at 3 mm only.

Finally, the inputs ‘‘R1 Fluxes’’ and ‘‘R2 fluxes’’ are associated (when needed) to the action EFF)

11.3.2 Actions or Outputs

To calibrate your data, you need to do the following actions leading to the output calibration, in order:

1. SELECT: Select the calibration parameters
2. PHCOR: Radiometric phase correction, equivalent to a ‘‘Monitor 0’’, see A.Dutrey lecture.
3. RF: Radio Frequency calibration
4. PHASE: Instrumental phase calibration versus time
5. EFF: Efficiency (Jy/K) calibration, to determine the flux densities of the amplitude calibrator. The inputs ‘‘R1 Fluxes’’ and ‘‘R2 Fluxes’’ should be used here.
6. AMP: Amplitude calibration versus time
7. PRINT: To produce the LaTeX and PostScript files containing the calibration curves (*e.g.* 28-feb-2001-x007.ps).

By typing GO, all the actions listed above are done sequentially. The reducer has just to type continue under Clic (or use the “Continue” button in the top left menu) at each step of the calibration process.

11.3.3 Results of the calibration

After a few general comments about quality of the data, including the measurement of the seeing (deduced from the rms of the fit of the temporal phase), the pages 1 and 2 summarize the calibration as follows:

- §1.1 The estimated flux densities of the calibrators at the observed frequencies.
- §1.2 The efficiencies of the antennas for Receiver 1 and 2 (which are deduced by fixing the flux of one or several calibrators).
- §1.3 The hour angle observed on the source
- §1.4 The table of the rms obtained on the RF calibration for receivers 1 and 2 in both upper and lower sidebands
- §1.5 The table of the rms obtained the temporal fit for the phase and the amplitude per baselines. Note that for receiver 2, the rms given in Col.1 is exactly the product of frequency ratio (rece 2/rece 1) times the rms obtained on receiver 1 because the phase on the receiver 1 is used to calibrate the phase on the receiver 2. After applying this phase correction, a second fit is performed on the residuals, its rms is displayed on Col.2. Col.3 gives the rms obtained on the amplitude (in %).

Chapter 12

UV Plane Analysis

Robert Lucas

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

12.1 *uv* tables

After calibration with CLIC, the calibrated data may be stored in a particular file called a '*uv* table'. This is useful because much of the data in the CLIC data file are not needed any more: atmospheric parameters, total powers, image side band visibilities, data from other receivers may be discarded at this stage. All that counts is: the data that are needed to describe the source itself, the sky frequency that was observed, ... One may for instance create a *uv* table for the continuum and one for each line that was observed.

These *uv* tables are just special GILDAS tables suited for *uv* data handling that are created by CLIC. Mapping consists of transforming these tables into something more meaningful for the astronomer, either images or numbers like positions, flux densities, sizes, etc. However a good part of the data evaluation and analysis can be directly performed on the *uv* data itself, before performing any of the complex operations involved in creating an image (Fourier transform and deconvolution). Direct analysis of the *uv* data is the subject of this Lecture.

12.1.1 *uv* table contents

A *uv* table is a file in the Gildas Data Format, of dimensions [$3N_{\mathrm{C}}$+7,N_{V}], for N_{C} spectral channels and N_{V} visibilities. The $3N_{\mathrm{C}}+7$ lines contain:

1. u in meters
2. v in meters
3. Scan number
4. Observation date (integer CLASSday number)
5. Time in seconds since above date

6. Number of start antenna of baseline
7. Number of end antenna of baseline
8. First frequency point (real part)
9. First frequency point (imaginary part)
10. First frequency point (weight)
11. Same for second frequency point, and so on

Thus for a given scan with N_A antennas, $N_\mathrm{A}(N_\mathrm{A}-1)/2$ visibilities are recorded.

The table header has the standard form of a GILDASImage. The header is available (for instance) by declaring:

```
GRAPHIC> SIC\DEFINE HEADER T co10.uvt READ
GRAPHIC> EXAMINE T%
```

For a table named `co10.uvt`. Some keywords convey a more precise meaning for *uv* tables:

`T%NDIM` should be 2

`T%DIM` contains $3N_\mathrm{C}+7$ and N_V

`T%RA,T%DEC` coordinates (radians) of the pointing center (the center of the primary beam).

`T%A0,T%D0` coordinates (radians) of the phase tracking center (a point source at this point should have zero phase); they are identical to RA and DEC when a table is first produced.

`T%EPOCH` The epoch of those coordinates. Should be 2000.0

`T%VELOFF, T%VELRES` The velocity of the reference channel, and the channel separation in velocity units (km/s)

`T%RESTFRE, T%FREQRES` The rest frequency, and the channel separation in frequency units (MHz)

`T%CONVERT[1,1]` the reference channel

`T%CONVERT[1,2]` the actual observing frequency at the reference channel (MHz); the one used to scale angular displacements from u, v coordinates in meters.

One may also examine directly the header by typing simply :

```
GRAPHIC> HEADER co10.uvt
```

12.1.2 How to create a *uv* Table

uv Tables are created by CLIC using the command `TABLE`.

A set of commands to create a *uv* table may look like:

```
! Reset the default options:
  SET DEFAULT
! find the useful scans:
  FILE IN 21-JAN-1998-H126
  SET SOURCE IRC+10216
  SET RECEIVER 1
  SET PROCEDURE CORRELATION
  SET QUALITY AVERAGE
  FIND
! calibration options:
```

```
  SET AMPLITUDE ANTENNA RELATIVE
  SET PHASE ANTENNA RELATIVE INTERNAL ATMOSPHERE
  SET RF ANTENNA ON
! table creation:
  SET SELECTION LINE LSB L01
  TABLE HCN NEW /FREQUENCY HCN  88631.85 /RESAMPLE 19 10 -27 2.12 V
```

All but the last two commands should be familiar at this point.

- The first new command, `SET SELECTION`, prepares the last one `TABLE`. It selects that the next table to be created will be a line table (i.e. with more than one spectral channel). The lower side band data will be used, and only the first subband of the correlator: `L01`.
- The last command `TABLE`, actually creates the table named `hcn.uvt`. The rest frequency, 88631.85 MHz, is set to be the reference used for the velocity scale. The data will be resampled to a velocity grid of 19 channels; the reference channel 10 will correspond to the LSR velocity -27 km/s; the channel spacing will be 2.12 km/s. Without `/RESAMPLE`, one would have got all the channels in the subband L01 with their original velocity separation. Without `/FREQUENCY`, the rest frequency present in the data (in the observing procedure) would have been used.

Using `/RESAMPLE`, one may avoid creating tables with too many channels (by discarding unused parts of correlator subbands) and choose the resolution that is actually needed.

If the data is spread on several files, one may go on by opening the other files, finding the data scans, and appending to the table:

```
  FILE IN 12-FEB-1998-H126
  FIND
  TABLE
  FILE IN 21-FEB-1998-H126
  FIND
  TABLE
...
```

(the arguments to `TABLE` need not to be repeated).

For continuum tables one may use:

```
SET SELECTION CONTINUUM DSB L01 TO L05 -
  /WINDOW 214405 214726 217476 217796 217837 217875
TABLE CONT-1MM NEW
```

Here we are using data from all the line subbands, but only in the three frequency windows: 214405 to 214726 MHz, 217476 to 217796MHz, and 217837 to 217875. This is of course to avoid the line emission of some molecules.

A standard menu is available under the CLIC main menu ("Create a UV Table"). After execution, a specific procedure is created to keep track of the options and parameters used. This procedure can subsequently be edited to add new data files (data files can also be added from the menu).

12.2 *uv* data plots

A procedure is available to do various plots from a continuum or line table. Its name is `UVALL` and it is called by clicking on "Interferometric UV operations" in the GRAPHIC standard menu. One has to select the first and last channel to be plotted (0 0 to get all channels) and the name of the parameters to be plotted in abscissa and ordinate. The following examples are the most useful plots:

uv **coverage:** to get an idea of the imaging quality that may be obtained, to check if one configuration has been forgotten, ...

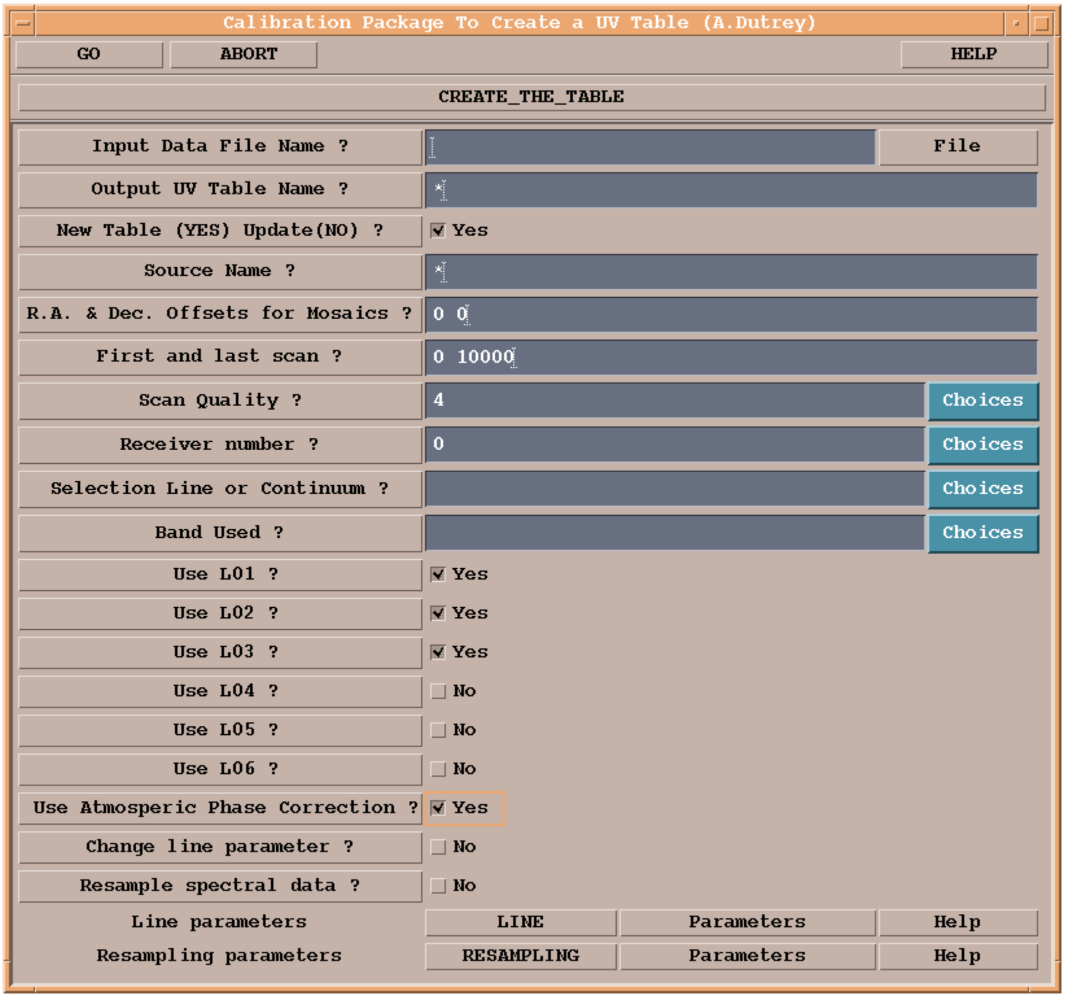


Figure 12.1: "Create a UV Table" menu in CLIC

weight vs. number: check if some data got strange weights (e.g., zero) for any reason

Amplitude vs. antenna spacing: quite useful if a source is strong to see if it looks resolved. Also check for spurious high amplitude points.

Amplitude vs. weight: another useful check: spurious high-amplitude points with non-negligible weight can cause a lot of harm in a map.

These plotting facilities are also implemented in the MAPPING program as a command (`SHOW UV`).

12.3 Data editing

Editing the interferometer data is not very flexible when done in the *uv* tables. If a problem occurs, is is not easily diagnosed since many of the parameters associated with the data acquisition are not present: atmospheric data, total powers, ... It is however useful when something strange occurs in the mapping process to do *uv* plots of the input table to look for the faulty data. Using time or scan number as X coordinate is recommended. One then may go back to the CLIC program and flag the faulty data, tag the corresponding scans, and keep a log of these problems so that they are not encountered again when the *uv* tables are reconstructed for any reason.

Two tasks have been written that may directly edit the data in *uv* tables:

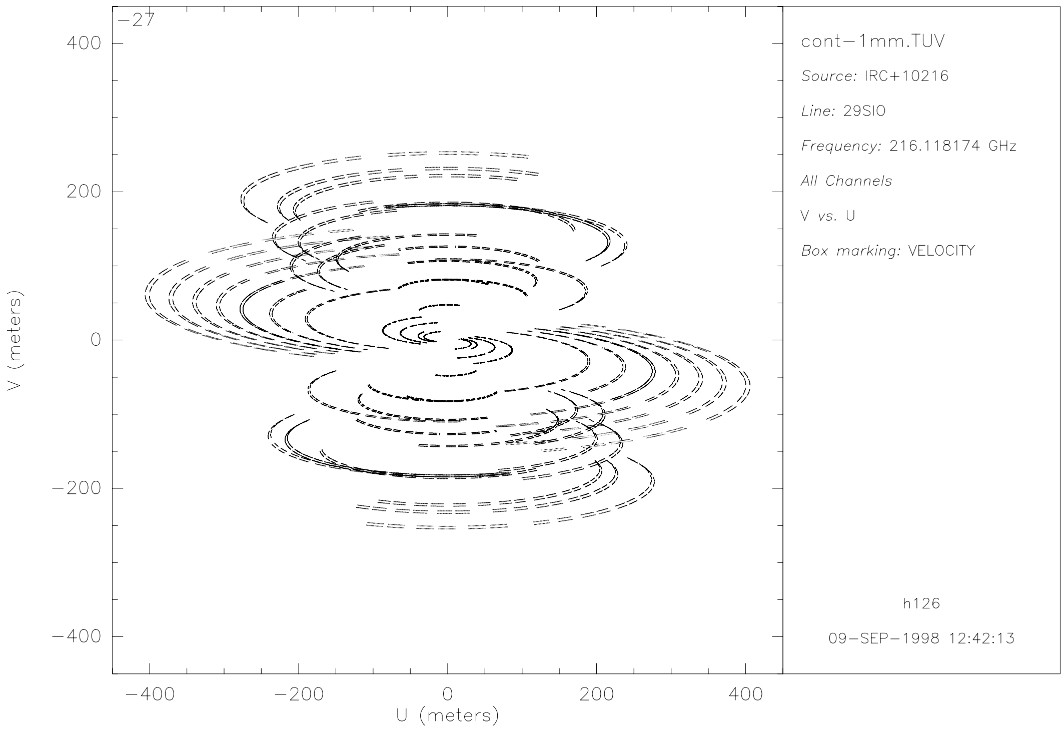


Figure 12.2: Example of a *uv* coverage plot

UV_CLIP flags all visibilities larger than a given flux: this deletes visibility points with totally wrong numbers, if any.

UV_FLAG deletes visibility points in a give time interval for a given baseline.

Both tasks work by setting the corresponding weight to zero: their action is irreversible (you will have to reconstruct the *uv* table to go back).

The MAPPING program provides a more efficient, simpler and reversible interactive tool to flag parts of a *uv* data set (command UV_FLAG).

12.4 Position shift

For the purpose of further data reduction it may be necessary to change the phase center of the *uv* data. This is done by a simple rotation of the phases (multiplication by a complex factor). One uses the task UV_SHIFT for this purpose.

12.5 Averaging

12.5.1 Data compression

One may simply wish to average several channels to increase the signal to noise ratio (use tasks UV_COMPRESS and UV_AVERAGE).

12.5.2 Circular averaging

For sources with circular symmetry it may be necessary to obtain the variation of amplitude with antenna spacing, in order to compare the amplitude data with models. For this purpose, with task UV_CIRCLE one

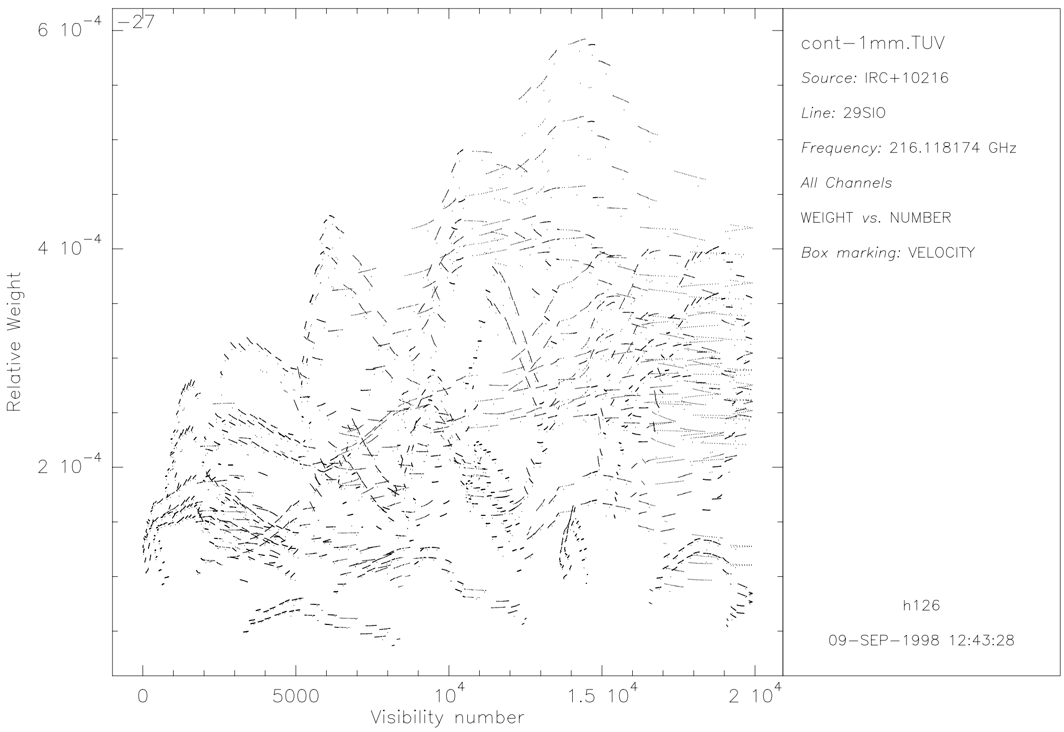


Figure 12.3: Weight versus visibility number plot

takes the mean of all the visibilities in concentric rings in the *uv* plane. The output has the format of a *uv* table (except that all *v*'s are zero), and may be plotted with `UVALL` (fig. 12.5).

Use this sort of averaging with caution: the phase center must accurately coincide with the source position or the amplitude of the visibility average will decrease on long spacings (use `UV_FIT` and `UV_SHIFT`. One may also do this kind of averaging in separate sectors in the *uv* plane, to check for asymmetries, provided the *uv* plane is well sampled (task `UV_CUTS`).

12.6 Model fitting

Model fitting is the oldest way of analyzing interferometer data. It was effectively used in the times where the coverage of the *uv* plane was too scarce to even think of creating an image by Fourier transform. One assumes a simple source model depending of a few parameters (source position, flux, size) and fits the visibility function of that model to the visibility data. Of course one may use a linear combination of several source models since the Fourier transform is linear. This is performed using the GILDAS task `UV_FIT`. The result may be displayed with the procedure `PLOTFIT`. Both are available in the panel "Interferometric UV operations" from the GRAPHIC standard menu.

Table 12.1 gives examples of a few models and their visibility functions. For source models with a circular symmetry, the visibility function is split into a radial dependent amplitude and a phase factor which depends only on the source position.

Some sources are actually so simple that this method may be used to a good accuracy (fig. 12.6).

Quite often this method is used for sources that are unresolved or not well resolved at a given frequency; for instance a SiO maser may consist of several point-source components at different velocities. Fitting a point source in each channel one derives a "spot map" (figs 12.7,12.8).

Parameters:		Variables:	
x_0	RA position	x, y	sky position
y_0	DEC position	r	$\sqrt{(x-x_0)^2+(y-y_0)^2}$
S	Source flux	u, v	projected spacing
b	HP size	q	$\sqrt{u^2+v^2}$

Name	Model	Visibility
Point source	$S\ \delta(x-x_0, y-y_0)$	$S\ e^{-2i\pi(ux_0+vy_0)}$
Gaussian	$\frac{4S}{\pi b^2 \log 2}\ e^{-4\log 2 \frac{r^2}{b^2}}$	$S\ e^{-\pi^2/4/\log 2\ (bq)^2}\ e^{-2i\pi(ux_0+vy_0)}$
Disk	$\frac{4S}{\pi b^2}$ where $\|r\| < b$	$SJ_1(\pi bq)\ e^{-2i\pi(ux_0+vy_0)}$

Table 12.1: A few source simple source models and their visibility functions

12.6.1 Position measurement

For a source with central symmetry the task `UV_CENTER` determines the source position by using only the phases. Alternatively the task `UV_FIT` may be used to fit the amplitudes and phases at the same time, or e.g. to simultaneously fit a pair of sources.

12.7 Continuum source subtraction

It is straightforward to subtract a point source at the phase center in the *uv* data: one simply subtract a real number (the source flux) from all the visibilities.

The task `UV_SUBTRACT` subtracts a time-averaged continuum *uv* table from a spectral line table (this assumes that the continuum and the line have been observed simultaneously), providing a new table with the line emission. Note that if the source is too complex, the time averaging (needed to avoid increasing the noise level in the resulting table), may affect the structure of the subtracted continuum image.

If the continuum data was not observed in the same session, or is known only from other sources, one may build a *uv* table of the continuum using the task `UV_MODEL`. This tasks computes that table from a model image or data cube; it computes the corresponding visibilities at *uv* coordinates taken from a reference *uv* table (e.g. the table out of which one wants to subtract the continuum model).

12.8 Self calibration by a point source

In some cases the continuum is a point source, and is strong enough to be used to reference the phases and amplitudes of the line data. The phase and amplitude referencing can be done in CLIC command `STORE PHASE /SELF` or in the *uv* tables using task `UV_ASCAL`.

For this one creates separately a line table and a continuum table. Naturally both must have been observed simultaneously. It is strongly advised to self-calibrate the phase first, then the amplitude using a longer smoothing time. `UV_ASCAL` allows in addition to subtract (part of) the input reference source, typically to provide continuum-free spectral line maps.

Note that this operation will work also if the source is somewhat extended, but with central symmetry; in that case only the phases can be self-calibrated.

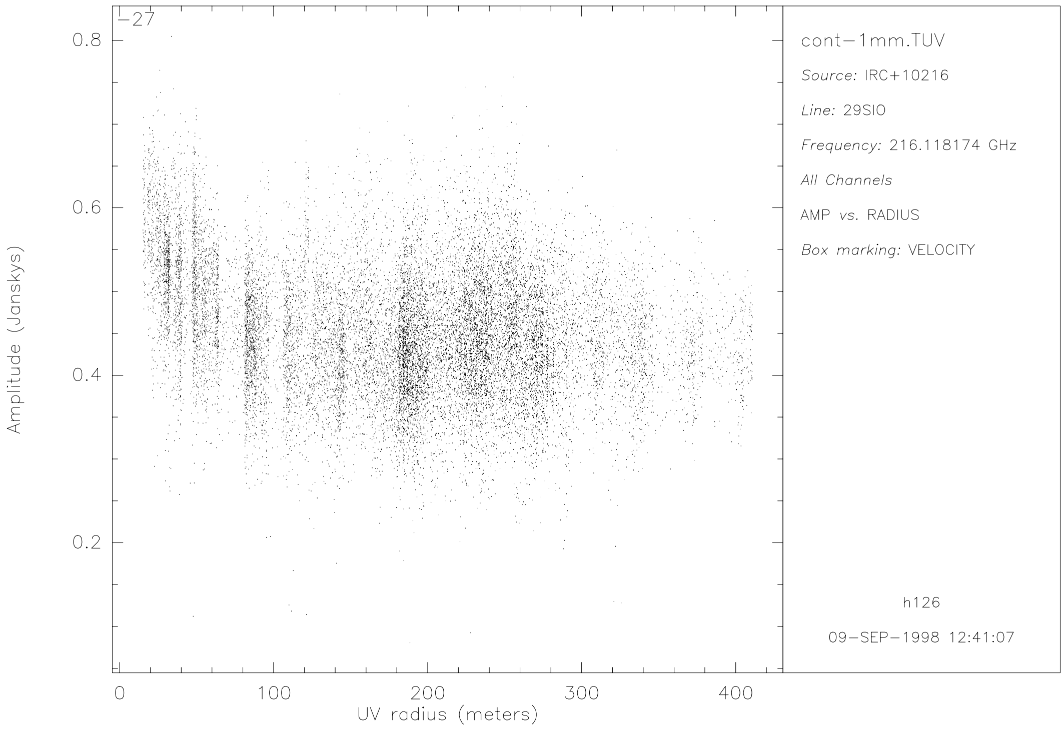


Figure 12.4: Amplitude versus antenna spacing plot

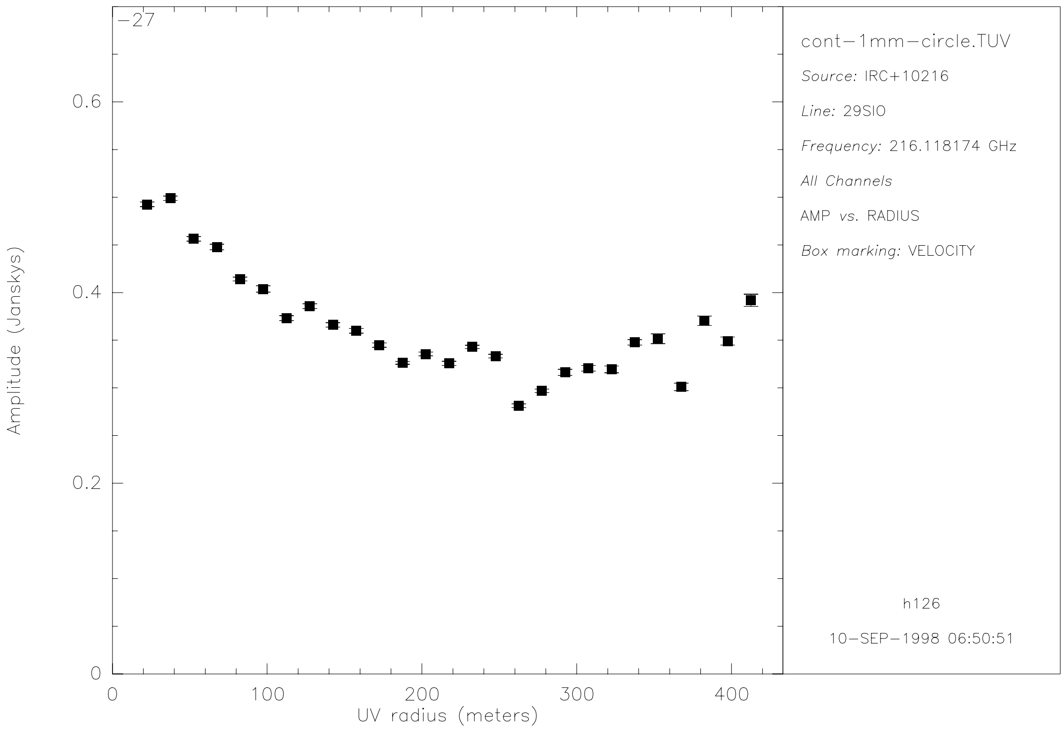


Figure 12.5: Example of a circular average plot (same data as fig. 12.4)

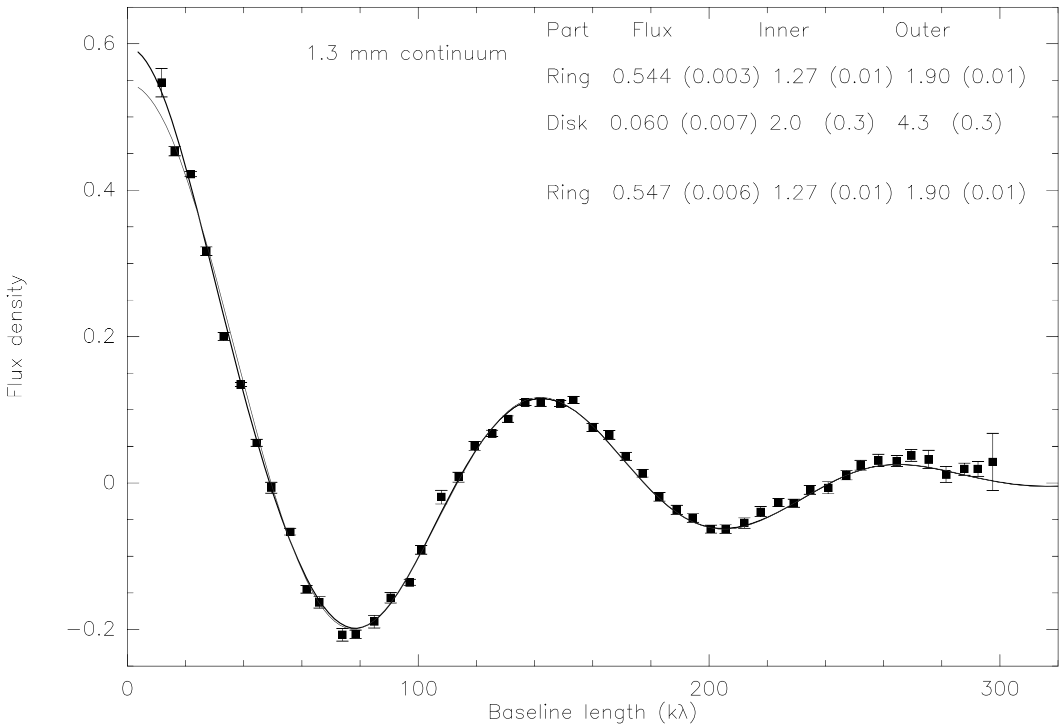


Figure 12.6: *uv*-plane fit to the disk around GG Tau at 1.3 mm (from [Guilloteau et al 1999]).

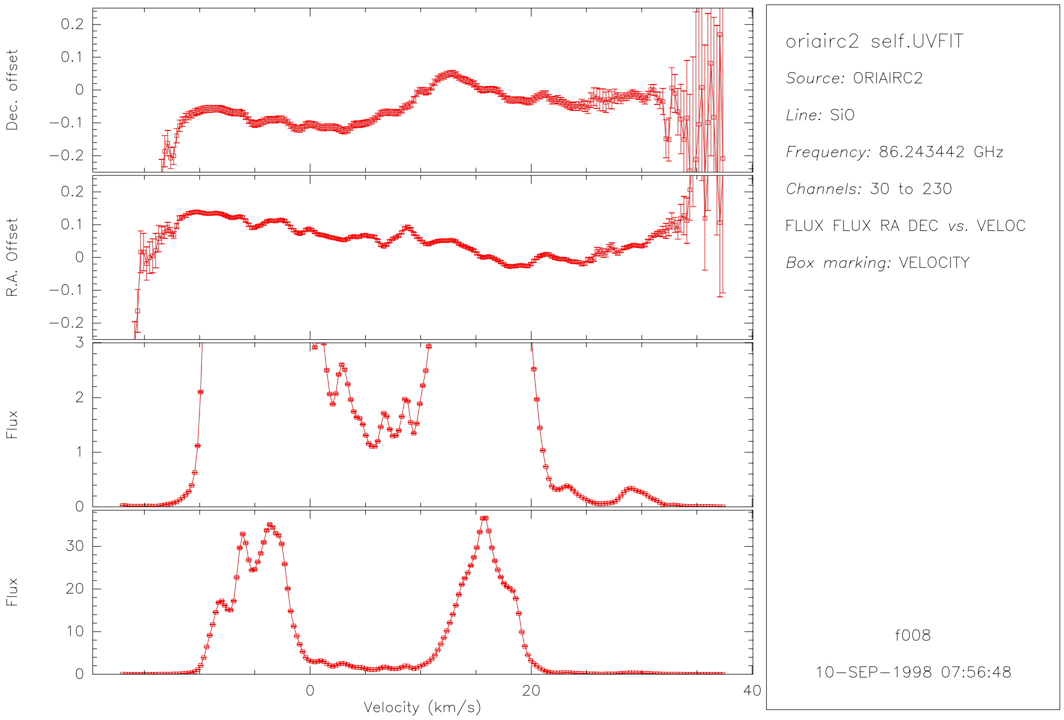


Figure 12.7: Result of a multi-channel point source fit to the Orion SiO (v=2, J=2-1) maser

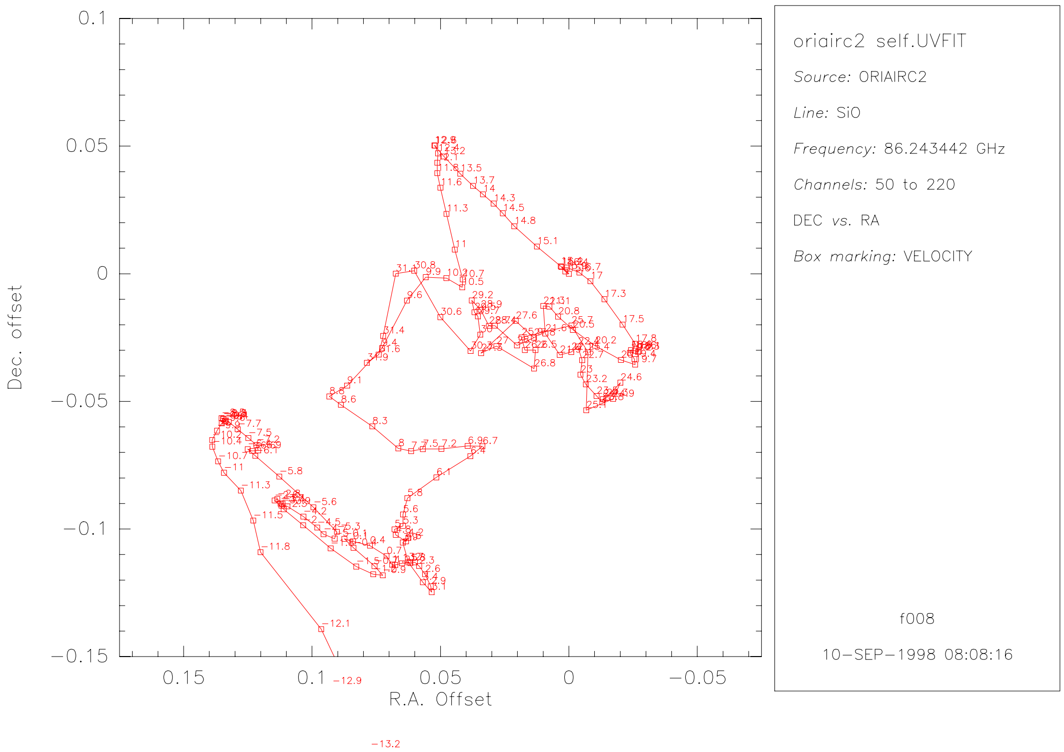


Figure 12.8: “Spot map” of the Orion SiO (v=2, J=2-1) maser

Chapter 13

The Imaging Principles

Stéphane Guilloteau

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

Assuming identical antennas, we have shown in previous lectures that an interferometer measures the *visibility function*

$$V(u,v) = \iint B(x,y)\, I(x,y)\, e^{-2i\pi(ux+vy)}\, dx\, dy \tag{13.1}$$

over an ensemble of points $(u_i, v_i), i = 1, n$, where $B(x,y)$ is the power pattern of the antennas and $I(x,y)$ the sky brightness distribution.

The imaging process consists in determining as best as possible the sky brightness $I(x,y)$. Since Eq.13.1 is a convolution, the imaging process will involve deconvolution techniques.

Let $S(u,v)$ be the sampling (or spectral sensitivity) function

$$\begin{aligned} S(u,v) \neq 0 &\Longleftrightarrow \exists i \in 1,n\; (u_i,v_i) = (u,v) \\ S(u,v) = 0 &\Longleftrightarrow \forall i \in 1,n\; (u_i,v_i) \neq (u,v) \end{aligned} \tag{13.2}$$

The spectral sensitivity function S contains information on the relative weights of each visibility, usually derived from theoretical noise.

Let us define

$$I_w(x,y) = \iint S(u,v)\, W(u,v)\, V(u,v)\, e^{2i\pi(ux+vy)}\, du\, dv \tag{13.3}$$

where $W(u,v)$ is an arbitrary *weighting function*. Since the Fourier Transform of a product of two functions is the convolution of the Fourier Transforms of the functions, $I_w(x,y)$ can be identified with

$$I_w(x,y) = (B(x,y) I(x,y)) ** (D_w(x,y)) \tag{13.4}$$

where

$$D_w(x,y) = \iint S(u,v)\, W(u,v)\, e^{2i\pi(ux+vy)}\, du\, dv = \widehat{SW} \tag{13.5}$$

$D_w(x,y)$ is called the *dirty beam*, and is directly dependent on the choice of the *weighting function* W, as well as on the spectral sensitivity function S. $I_w(x,y)$ is usually called the dirty image.

Fourier Transform, which allows to directly derive I_w from the measured visibilities V and spectral sensitivity function S, and **Deconvolution**, which allows to derive the sky brightness I from I_w, are thus two key issues in imaging (see Eq.13.4).

13.1 Fourier Transform

The first step in imaging is to evaluate the dirty image, using Fourier Transform. Several techniques are available.

13.1.1 Direct Fourier Transform

The simplest approach would be to directly compute sin and cos functions in Eq.13.4 for all combinations of visibilities and pixels in the image. This is straightforward, but slow. For typical data set from the VLA, which contain up to 10^5 visibilities per hour and usually require large images (1024×1024 pixels), the computation time can be prohibitive. On the other hand, the IRAM Plateau de Bure interferometer produces about 10^4 visibilities per synthesis, and only require small images (128×128). The Direct Fourier Transform approach could actually be efficient on vector computers for spectral line data from Plateau de Bure interferometer, because the sin and cos functions needs to be evaluated only once for all channels. Moreover, the method is well suited to real-time display, since the dirty image can be easily updated for each new visibility.

13.1.2 Fast Fourier Transform

In practice, everybody uses the Fast Fourier Transform because of its definite speed advantage. The drawback of the methods is the need to regrid the visibilities (which are measured at arbitrary points in the (u,v) plane) on a regular grid to be able to perform a 2-D FFT. This gridding process will introduce some distortion in the dirty image and dirty beams, which should be corrected afterwards. Moreover, the gridded visibilities are sampled on a finite ensemble. As discussed in more details below, this sampling introduces aliasing of the dirty image (and beam) in the map plane.

13.1.3 Gridding Process

The goal of gridding is to resample the visibilities on a regular grid for subsequent use of the FFT. At each grid point, gridding involves some sort of interpolation from the values of the nearest visibilities. The visibilities being affected by noise, the interpolating function needs not fit exactly the original data points. Although any interpolation scheme could a priori be used, such as smoothing spline functions, it is customary to use a *convolution* technique to perform the gridding.

Using a convolution is justified by several arguments. First, from Eq.13.1, $V = \widehat{BI} = \hat{B} ** \hat{I}$. Hence V is already a convolution of a (nearly Gaussian) function $\hat{B}$ with the Fourier Transform of I. Nearby visibilities are not independent. Second, as mentioned above, exact interpolation not desirable, since original data points are *noisy samples of a smooth function.* Third, if the width of the convolution kernel used in gridding is small compared to $\hat{B}$, the convolution added in the gridding process will not significantly degrade the information. Last, but not least, it is actually possible to correct for the effects of the convolution gridding.

To demonstrate that, let G be the gridding convolution kernel. Eq.13.3 becomes

$$I_w^g \rightleftharpoons G ** (S \times W \times V) \tag{13.6}$$

$$I_w^g \rightleftharpoons G ** (S \times W \times V) \quad I_w^g = \hat{G} \times (\widehat{SW} ** \hat{V}) \quad = \hat{G} \times I_w \tag{13.7}$$

$$D_w^g \rightleftharpoons G ** (S \times W) \qquad D_w^g = \hat{G} \times \widehat{SW} \tag{13.8}$$

$$\frac{I_w^g}{\hat{G}} = \frac{D_w^g}{\hat{G}} ** (BI) \tag{13.9}$$

Thus the dirty image and dirty beams are obtained by dividing the Fourier Transform of the gridded data by the Fourier Transform of the gridding function.

13.2 Sampling & Aliasing

Sampling on a regular grid is equivalent to multiplying by a series of periodically spaced delta functions, or the so-called *shah* function:

$$[\frac{1}{\Delta u}]\mathrm{III}(\frac{u}{\Delta u}) = \sum_{k=-\infty}^{\infty} \delta(u - k\Delta u) \quad (13.10)$$

The Fourier Transform of the *shah* function above is the *shah* function

$$\mathrm{III}(x\Delta u) = \frac{1}{\Delta u} \sum_{m=-\infty}^{\infty} \delta(x - \frac{m}{\Delta u}) \quad (13.11)$$

Hence, sampling the visibilities V results in convolving its Fourier Transform $\hat{V}$ by a periodic *shah* function. This convolution reproduces in a periodic way the Fourier Transform of the visibilities $\hat{V}$.

If the Fourier Transform of the visibilities $\hat{V}$, i.e. the brightness distribution BI, has finite support ΔX, the replication poses no problem provided the support is smaller than the periodicity of the *shah* function, i.e.

$$(\Delta u)^{-1} \geq (\Delta X) \qquad \Delta u \leq (\Delta X)^{-1} \quad (13.12)$$

If not, data outside $(\Delta u)^{-1}$ are *aliased* in the *imaged area* $(\Delta u)^{-1}$.

In aperture synthesis, finite support is ensured to first order by the finite width of the antenna primary beam B. However, strong sources in the antenna sidelobes may be aliased if the imaged area is too small. Moreover, the **noise** does not have finite support. White noise in the *uv* plane would result in white noise in the map plane. In practice, the noise in the *uv* plane is not completely white. However, it is support limited (since only a finite region of the *uv* plane is sampled in any experiment). Accordingly, its Fourier Transform in the map plane is not support limited. Noise aliasing thus occurs, and produces an increased noise level at the map edges.

13.3 Convolution and Aliasing

The combination of Gridding and Sampling produces the *uv* data set

$$\begin{aligned} V_m &= \frac{1}{\Delta u \Delta v}\mathrm{III}(\frac{u}{\Delta u}, \frac{v}{\Delta v}) \times (G **(S \times W \times V))(u,v) & (13.13)\\ &= \mathrm{III} \times (G **(S \times W \times V))/(\Delta u \Delta v) & (13.14)\end{aligned}$$

which analogous with Eq.13.6

The Fourier Transform of this *uv* data set is

$$\begin{aligned} \hat{V_m} &= \mathrm{III}(x\Delta u, y\Delta v) **(\hat{G} \times (\widehat{SW} \times \hat{V} & (13.15)\\ &= \mathrm{III} **(\hat{G} \times (\widehat{SW} **(BI))) & (13.16)\end{aligned}$$

V_m is thus the sky brightness multiplied by the primary beam (BI), convolved by the the dirty beam $\widehat{SW}$, then multiplied by the Fourier transform of the gridding function $\hat{G}$ and periodically replicated (by the convolution with last Shah function).

Accordingly, aliasing of $\hat{G}$ in the map domain will thus occur. Note at this stage that, providing aliasing of $\hat{G}$ remains negligible, an exact convolution equation is preserved

$$\frac{\hat{V_m}}{\hat{G}} = \widehat{SW} **BI \quad (13.17)$$

The gridding function will thus have to be selected to minimize aliasing of $\hat{G}$. This criterion will depend on the image fidelity required. Obviously, if the data is very noisy, aliasing of the $\hat{G}$ can be completely negligible.

Furthermore, the weighting function W is usually smooth, while the gridding function G is a relatively sharp function (since it ensures the re-gridding by convolution from *nearby* data points). Thus, to first order $G **W = W$, and we could rewrite Eq.13.14 as

$$V_m \quad = \mathrm{III} \times W \times (G **(S \times V))/(\Delta u \Delta v) \tag{13.18}$$

Hence, the weighting can be performed *after* the gridding. The choice of weighting before or after gridding is essentially based on computational speed or algorithmic simplicity.

Let us focus on the choice of the gridding function. The gridding function will be selected according to the following principles:

1. small support, typically one or two cells wide (Δu).
2. small aliasing.
3. fast computation.

Points 1 and 2 are contradictory, since a small support for G implies a large extent of $\hat{G}$. Some compromise is required. For simplicity, gridding functions are usually selected among those with separable variables:

$$G(u, v) = G_1(u) G_1(v)$$

although this breaks the rotation invariance.

The simplest gridding function is the **Rectangular function**

$$G(u) \quad = \quad \frac{1}{\Delta u} \Pi(\frac{u}{\Delta u}) \tag{13.19}$$

$$\hat{G}(x) \quad = \quad \frac{sin(\pi \Delta u x)}{\pi \Delta u x} \tag{13.20}$$

where Π is the unit rectangle function. Obviously, aliasing will be important, since the `sinc` function falls off very slowly.

A better choice could be the **Gaussian function**

$$G(u) \quad = \quad \frac{1}{\alpha \Delta u \sqrt{\pi}} e^{-(u/\alpha \Delta u)^2} \tag{13.21}$$

$$\hat{G}(x) \quad = \quad e^{-(\pi \alpha x \Delta u)^2} \tag{13.22}$$

By proper selection of α (not too small, not too large), a compromise between computation speed (better for small α) and aliasing (better for large α) can be found. $\alpha = 2\sqrt{ln(4)} \simeq 0.750$ is a standard choice.

However, a Gaussian still has fairly significant wings. $\hat{G}$ should ideally be a rectangular function (1 inside the map, 0 outside). G would be a `sinc` function, but this falls off too slowly, and would require a lot of computations in the gridding. Moreover, the (unavoidable) truncation of G would destroy the sharp edges of $\hat{G}$ anyhow. Hence the idea to use an apodized version of the `sinc` function, the **Gaussian-Sinc function**

$$G(u) \quad = \quad \frac{\sin \pi u/(\alpha \Delta u)}{\pi u} e^{-(u/(\beta \Delta u))^2} \tag{13.23}$$

$$\hat{G}(x) \quad = \quad \Pi(\alpha x \Delta u) * (\sqrt{\pi} \beta \Delta u e^{-(\pi \beta x \Delta u)^2}) \tag{13.24}$$

It provides good performance for $\alpha = 1.55$ and $\beta = 2.52$.

The empirical approaches mentioned above do not guarantee any optimal choice of the gridding function. A completely different approach is based on the desired properties of the gridding function. We actually want $\hat{G}$ to fall off as quickly as possible, but G to be support limited. Mathematically, this defines a class of functions known as **Spheroidal functions**. Spheroidal functions are solutions of differential equations, and cannot be expressed analytically. In practice, this is not a severe limitation since numerical representations can be obtained by tabulating the gridding function values. Given the limited numerical accuracy of the computations, the tabulation does not require a prohibitively fine sampling of the gridding function, and is quite practical both in term of memory usage and computation speed. Tabulated values are used in the task `UV_MAP`.

Note that the finite accuracy of the computation may ultimately limit the image dynamic range.

13.4 Error Analysis

We thus succeeded to preserve a convolution equation, with the slight restrictions due to the aliasing and gridding correction. Let us explore now what *errors* or *systematic* effects may appear in the image plane.

First, consider the **noise**. Aliasing increases the noise level at the map edges (by noise aliasing and then by the gridding correction since this amounts to divide by the Fourier Transform of the gridding function, which is unity at the map center, but smaller at the map edges). For example, the noise increases by a factor $(\pi/2)^2$ at map corners for the Gaussian-Sinc function. Near the map center, the effect is negligible. Note that for a given field of view, the noise increase can be arbitrarily limited by making a sufficiently large image, but this has a high computational price.

Concerning errors, it is important to separate two main classes of errors.

Additive errors

The Fourier transform being linear, additive errors result in artificial structure *added* to the true map, e.g.
- A single spurious visibility will produce fringes in the map
- An additive real term (correlator offset), will produce a point source at the phase tracking center.

Multiplicative errors

A multiplicative term on the visibility distorts the image, since

$$V(u,v) \times \epsilon(u,v) \longleftrightarrow \hat{V}(x,y) ** \hat{\epsilon}(x,y)$$

i.e. the map is convolved by the Fourier transform of the error. Calibration errors (in phase or amplitude) are of this type. Among these, the **seeing** should not be neglected.

Phase calibration errors result in antisymmetric patterns.

13.5 Weighing and Tapering

There is still a free parameter in the image construction process: the weighting function. At uv table creation, the sampling function is defined as

$$S(u,v) = \frac{1}{\sigma^2(u,v)} \tag{13.25}$$

where the noise σ is computed from system temperature, bandwidth, integration time, and system efficiency (including quantization and decorrelation).

$$\sigma(u,v) = \frac{\mathcal{J}_I T_{\mathrm{sys}}}{\eta_{\mathrm{Q}}\sqrt{2\Delta\nu t_{\mathrm{int}}}} \tag{13.26}$$

$\mathcal{J}_I$ being the antenna temperature to flux density conversion factor:

$$\mathcal{J}_I = \frac{2k}{\eta_{\mathrm{A}} A} \tag{13.27}$$

The weights $W(u,v)$ can be freely chosen. The selecting of weights is usually decomposed in two slightly different processes, called Weighting and Tapering.

Weighting deals with the local variations of weights for each grid cell after the gridding process. Since the original uv coverage is an ensemble of ellipses, the gridding may leave a weight distribution with very large dispersion. Weighting can be applied to uniformize this distribution.

On the other hand, **Tapering** consists in apodizing the uv coverage by $T(u,v) = exp(-(u^2+v^2)/t^2))$ where t is a tapering distance. This corresponds to smoothing the data in the map plane (by convolution with a gaussian).

The simplest possibility, called **Natural Weighting, without taper** is to keep the original spectral sensitivity function by setting $W(u,v) = 1$. This can be demonstrated to maximize sensitivity to *point sources* (i.e. sources smaller than the synthesized beam). Proper design (and use) of the array can ensure that the resulting synthesized beam is appropriate, in terms of size (angular resolution matched to the scientific goal) and shape (lowest possible sidelobes).

If the sources of interest are somewhat extended, tapering can be used to increase **brightness** sensitivity. Tapering also has the advantage of lowering the sidelobes. However, tapering is always throwing out some information, namely the long baselines part of the data set. Hence, it should be used either with moderate tapers, or as a complementary view on a data set. To increase brightness sensitivity, one should use preferentially compact arrays rather than tapering.

Uniform Weighting consists in selecting the weights $W(u,v)$ so that the sum of weights $\sum W \times S$ over a uv cell is a constant function (or zero if no uv data exists in that cell). The size (radius) of the uv cell is an arbitrary parameter. It can be the cell size resulting from the gridding process, i.e. the inverse of the field of view, but any other choice is possible. Using half of the dish diameter is well justified, since the visibilities are convolution of Fourier transform of the sky brightness by the Fourier transform of the primary beam. **Uniform Weighting** gives more weight to long baselines than natural weighting (because you spend less time per uv cell on long baselines than on short baselines for earth synthesis). **Uniform Weighting** produces smaller beam. Because it fills the uv plane more regularly, Uniform weighting could be thought also to produce lower sidelobes. However, because of the discontinuity of the weights at the edge of the sampled portion of the uv plane, the inner sidelobes tend to be increased, unless some tapering is combined with Uniform weighting.

Robust Weighting is a variant of uniform weighting which avoids to give too much weight to a uv cell with low natural weight. There are several ways to implement such a scheme. Roughly speaking, if the sum of natural weights in a cell is less than a threshold, the weighting is unchanged, if it is more, the weight is set to this threshold. Let S_n be the natural weight of a cell, and S_t a threshold for such weight. Robust weighting could be implemented by selecting the weight W as

$$\begin{aligned} S_n < S_t &\rightleftharpoons W = 1 \\ S_n > S_t &\rightleftharpoons W = S_t/S_n \end{aligned} \tag{13.28}$$

or a more continuous formula like

$$W = \frac{1}{\sqrt{1+S_n^2/S_t^2}} \tag{13.29}$$

Robust weighting combines the advantages of **Natural** and **Uniform** weighting, by increasing the resolution and lowering the sidelobes without degrading too much the sensitivity. By adjusting the threshold, it approaches either case (large threshold $\longleftrightarrow$ Natural, small threshold $\longleftrightarrow$ Uniform).

Weighting and Tapering reduce point source sensitivity by

$$\sqrt{\sum T^2W^2}/(\sum TW) \tag{13.30}$$

13.6 The GILDAS implementation

We have now introduced the basic parameters of the imaging process: gridding, weighting and tapering. The main imaging task in the GILDAS software is UV_MAP. Before using UV_MAP, it is also recommended to use the associated task UV_STAT which evaluates the beam sizes, point source and brightness sensitivity as function of taper or robust weighting parameter.

Although the choice of configurations for the Plateau de Bure interferometer has been performed in order to optimize the uv coverage for most observing conditions, robust weighting can often offer a better compromise, unless signal to noise is insufficient. Task UV_STAT also suggests appropriate pixel sizes for UV_MAP

The imaging task UV_MAP is controlled by the following parameters:

- `MAP_SIZE`
  The number of pixels in each direction. This should be powers of 2.

- `MAP_CELL`
  The pixel size, in arcsecond, in each direction. It should respect proper sampling compared to the synthesized beam width. In practice, 3 – 4 pixels per beam width are required. Task `UV_STAT` can compute the optimum value for this parameter. Note that the imaged area is `MAP_SIZE` × `MAP_CELL`

- `MCOL`
  For spectral line data, the first and last channel to be imaged. (0,0) means all data.

- `WCOL`
  The channel from which the natural weights S are taken. `UV_MAP` produces only one beam for all channels (by default, there is an alternate option for experts). `WCOL = 0` is equivalent to `WCOL = (MCOL[1]+MCOL[2])/2`.

- `WEIGHT_MODE`
  `UN` for **Uniform** or `NA` for **Natural** weighting. `UN`iform weighting is actually a **Robust** weighting in `UV_MAP`.

- `UV_CELL`
  When UNiform weighting is used, `UV_CELL[1]` is the UV cell diameter (in meters), and `UV_CELL[2]` is the threshold for robust weighting: 1 corresponds to the mean natural weight of all cells. `UV_CELL[1]` should normally be 7.5 m for Plateau de Bure data.

- `CONVOLUTION`
  This is the convolution type for gridding. Choices are offered for test purposes, but `CONVOLUTION = 5` (Spheroidal) gives best results.

The other parameters are used to re-center the map (by phase shifting the *uv* data before imaging) when needed. This is convenient for **Mosaics**. `UV_MAP` performs all the imaging steps presented before: gridding, weighting, tapering, correction for gridding function, and computes the dirty beam and dirty image.

Both `UV_STAT` and `UV_MAP` are implemented as commands in the Mapping program, or as tasks available from the Graphic program. Using one or the other is a matter of personal preference.

13.7 Deconvolution

The first imaging step presented before leads to a convolution equation whose solution is the convolution product of the sky brightness distribution (apodized by the interferometer primary beam) by the dirty beam.

To derive the astronomically meaningful result, i.e. ideally the sky brightness, a deconvolution is required. Deconvolution is always a non linear process, and requires (in one way or another) to impose some constraints on the solution, or in other words to add some information, to better select plausible solutions. Such additional constraints can be explicit (e.g. positivity, or user specified finite support) or qualitative.

13.7.1 The CLEAN method

The standard deconvolution technique, Clean relies on such a qualitative constraint: it assumes that the sky brightness is essentially a ensemble of point sources (the sky is dark, but full of stars). The algorithm which derives from such an assumption is straightforward. It is a simple "matching pursuit"

1. Initialize a *Residual* map to the *Dirty* map

2. Initialize a *Clean component list* to zero.

3. Assume strongest feature in *Residual* map originates from a point source
4. Add a fraction γ (the *Loop Gain*) of this point source to the *Clean component list*, remove the same fraction, convolved with the dirty beam, from the *Residual* map.
5. If the strongest feature in the *Residual* map is larger than some threshold, go back to point 3 (each such step is called an iteration).
6. If the strongest feature is below threshold, of if the number of iterations $N_{\rm iter}$ is too large, go to point 7.
7. Convolve the *Clean component list* by a properly chosen *Clean Beam* (this is called the restoration step).
8. add to the result the *Residual* map to obtain the *Clean Map*.

The CLEAN algorithm as a number of free parameters. The loop gain controls the convergence of the method. In theory, $0 < \gamma < 2$, but in practice one should use $\gamma \simeq 0.1 - 0.2$, depending on sidelobe levels, source structure and dynamic range. While high values of γ would in principle give faster convergence, since the remaining flux is $\propto (1-\gamma)^{N_{\rm iter}}$ if the object is made of a single point source, deviations from an ideal convolution equation force to use significantly lower values in order to avoid non linear amplifications of errors. Such deviations from the ideal convolution equation are unavoidable because of thermal noise, and also of phase and amplitude errors which distort the dirty beam.

The threshold for convergence and number of iterations define to which accuracy the deconvolution proceeds. It is common practice to CLEAN down to about the noise level or slightly below. However, in case of strong sources, the residuals may be dominated by dynamic range limitations.

The clean beam used in the restoration step plays an important role. It is usually selected as a 2-D Gaussian, which allows the convolution to be computed by a simple Fourier transform, although other choices could be possible. The size of the clean beam is a key parameter. It should be selected to match the (inner part of) the dirty beam, otherwise the flux density estimates may be incorrect. To understand this problem, let us note first that the units of the dirty image are undefined. Simply, a 1 Jy isolated point source appears with a peak value of 1 in the dirty map. This is no longer true (because of sidelobes) if there is more than one point source, or a fortiori, an extended source. The unit of the clean image is well defined: it is Jy per beam, which can easily be converted to brightness temperature from the effective clean beam solid angle and the observing wavelength. Now, assume the source being observed is just composed of 2 separate point sources of equal flux, and that the dirty beam is essentially a Gaussian. Let us clean the dirty image in such a way that only 1 of the 2 point sources is actually included in the clean component list. If we restore the clean image with a clean beam which is, e.g. twice smaller than the original dirty beam, the final result will undoubtedly be odd. The second source would appear extended and have a larger flux than the first one. No such problem appears if the clean beam matches the dirty beam. Admittedly, the above example shows a problem which results from a combination of two effects: an inappropriate choice for the clean beam, and an insufficient deconvolution. However, the second problem always exists to some extent, because of noise in the original data set. Hence, to minimize errors, it is important to match the clean and dirty beams.

Note that in some circumstances, there may be no proper choice. An example is a dirty beam with narrow central peak on top of a broad "shoulder". Small scale structures will be properly reconstructed, but larger ones not.

The last step in the CLEAN method plays a double role. On one hand, it protects against insufficient deconvolution. Furthermore, since the residual image should be essentially noise if the deconvolution has converged, it allows noise estimate on the cleaned image.

13.7.2 Interpretation of CLEAN

If CLEAN converged, the *Clean component list* is a plausible solution of the measurement equation (within the noise), but it is not unique... Hence, because of convolution by the clean beam, the clean image is **not** a solution. However, besides allowing a reasonable definition of the image unit in case of incomplete convergence, there are two reasons to convolve by a clean beam. First, convolution by the clean beam smears out artifacts due to extrapolation beyond the measured area of the *uv* plane. This is an **a posteriori regularization**. Second, the clean components are forced to reside on the grid defined by the image. This discrete representation has a number of limitations (e.g. necessity of negative clean components, limited accuracy due to the finite size of the component list), which are reduced by convolution by the clean beam, because the clean image then has finite resolution and can be properly represented on a discrete grid provided the Nyquist sampling is preserved.

An important property of CLEAN is that (to first order) only the inner quarter of the dirty image can be properly cleaned. This is easily understood when dirty beam and dirty images are computed on the same grid size, since a source at one edge of the inner quarter requires knowledge of the dirty beam sidelobes beyond the map size to be deconvolved from the opposite edge. However, this also remains true if one computes the dirty beam on a twice larger grid than the dirty image: more than the inner quarter can be deconvolved, but because of aliasing, the map edges can never be.

Finally, CLEAN offers a very simple way to impose further constraints on the class of solution which is acceptable, by allowing definition of a *support*. This can be the standard (simple or multiple) *Clean Box* available in many non interactive implementations, or a user defined mask in interactive implementations. The search region can even be modified from iteration to iteration to help clean convergence. Such a flexible support is available inside the MAPPING program. Note however that the Clean Box or support should not be too limited: cleaning the noise is necessary too (as well as incorporating negative Clean component).

13.7.3 The CLEAN variants

The original CLEAN method is due to [Hogbom 1974]. Several variants exist.

One of the most popular (CLARK) is due to [Clark 1980], and involves minor and major cycles. In **Minor** cycles, an Hogbom CLEAN is performed, but with a truncated dirty beam, and only on the list of brightest pixels. This search is fast, because of the dirty beam truncation and because of the limited support. The Clean components identified during the minor cycles are removed at once by a FFT during a **Major** cycle. Because removal is done by FFT, slightly more than the inner map quarter can be cleaned.

A second variant, called MX, due to [Cotton & Schwab 1984], is similar to the CLARK method, except that the Clean components are removed from the *uv* table at the **Major** cycle stage (and thus the imaging process is repeated at each major cycle). This avoid aliasing of sidelobes, allows to clean more than the inner quarter, but is relatively slow because of the re-imaging at major cycles. Unless disk storage is a real problem, a faster result of equal is obtained by standard Clean with a twice larger map.

The next variant, called SDI (from [Steer et al 1984]), is again like the CLARK method, but in **Minor** cycles, no deconvolution is performed, but only a selection of the strongest components down to some threshold. **Major** cycles are identical to those of the CLARK method. Although the principle is simple, the implementation is not easy because of normalization subtleties in the minor cycle stage. This method is reasonably well suited for more extended structures, but could become unstable if the threshold is inappropriate.

The Multi Resolution Clean (MRC, [Wakker & Schwartz 1988]) separates the problem in a smooth map and a difference map. Since the measurement equation is linear, both maps can be Cleaned (with Hogbom or Clark method) independently. This is faster than the standard CLEAN because the smooth map can be compressed by pixel averaging, and only fine structure left in difference map, so fewer Clean components are required.

13.7.4 The GILDAS implementation

All the above variants are implemented in the GILDAS software. All of them, except MX, are implemented both as tasks and as interactive commands in the MAPPING program. The later implementation allows definition of a flexible support constraint. The default method is CLARK. SDI & MRC are usually not necessary for Plateau de Bure, because of the small ratio between the field of view (primary beam) and the resolution (< 30).

MX is implemented only as a task, and not recommended because of its relatively slow speed. Since Plateau de Bure images are relatively small (128×128), it is easier to use a standard clean on larger images.

The GILDAS software does not include any implementation of the Maximum Entropy Method, MEM. The main reason is that MEM is not suited for limited uv coverage. But MEM also has some undesirable properties, among which its attempt to give a unique solution, with no physical justification, the noise dependent resolution, and the definition of a global criterium for adjustment to data. Furthermore, no noise estimate is possible on MEM deconvolved images.

Chapter 14

Advanced Imaging Methods: WIPE

Eric Anterrieu

Observatoire Midi-Pyrénées – UMR5572, 14, avenue Edouard Belin, F-31400 Toulouse

14.1 Introduction

This lecture is the second part of a series describing how the visibility samples provided by an interferometric device can be used to produce a high quality image of the sky.

WIPE is a regularized Fourier synthesis method recently developed in radio imaging and optical interferometry. The name of WIPE is associated with that of CLEAN, the well-known deconvolution method presented in the previous lecture, and intensively used by astronomers at IRAM as well as in many institutes, worldwide.

The regularization principle of WIPE refers to the Shannon sampling formula and to theoretical considerations related to multiresolution analysis. The notions of field and resolution appear via the definition of two key spaces: the *object space* and the *object representation space* (a subspace of the first). The complex visibilities define a function in another space: the *data space*. The functions lying in this space take their values on a frequency list which is the concatenation of the *experimental frequency list* and a *regularization frequency list*. The latter defines a virtual frequency coverage beyond the *frequency coverage to be synthesized*, up to the highest frequencies of the *scaling functions* generating the *object space*. This virtual sampling is performed at the Shannon rate corresponding to the synthesized field. The reconstructed image, also called the *neat map*, is defined as the function minimizing a regularized objective functional in which the data are damped appropriately. To describe WIPE we adopt a terminology derived from that of CLEAN.

In this lecture, we present the basic foundations of WIPE, and its implementation in the IRAM data processing software. The reader interested in the theoretical aspects and developments of WIPE is invited to consult the articles [Lannes et al 1994], [Lannes et al 1996], [Lannes et al 1997].

14.2 Object space

In the problems of Fourier synthesis encountered in astronomy, the *object function* of interest, Φ_o, is a real-valued function of an angular position variable $\sigma \equiv \mathbf{x} = (x, y)$. The geometrical elements under consideration are presented in Fig. 14.1.

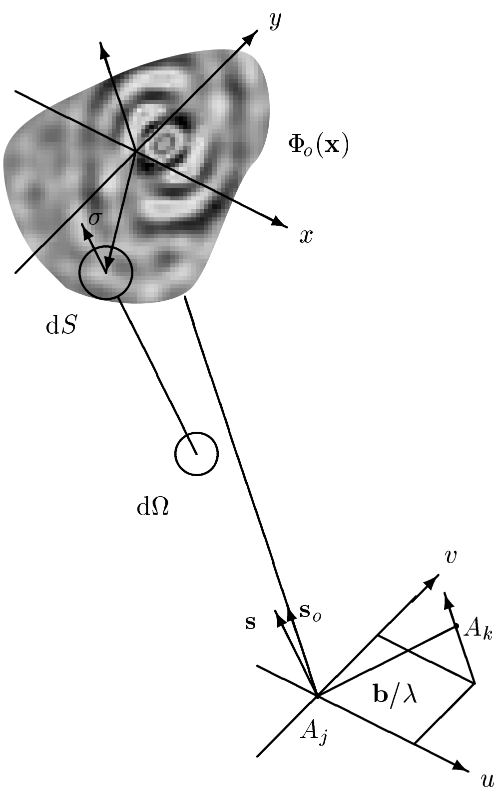


Figure 14.1: Traditional coordinate systems used to express the relation between the complex visibilities and the brightness distribution of a source under observation. Here, the two antennas A_j and A_k point toward a distant radio source in a direction indicated by the unit vector $\mathbf{s}$, and $\mathbf{b}$ is the interferometer baseline vector. The position pointed by the unit vector $\mathbf{s}_o$ is commonly referred to as the *phase tracking center* or *phase reference position*: $\mathbf{s} - \mathbf{s}_o = \sigma$.

The *object model variable* ϕ lies in some *object space* H_o whose vectors, the functions ϕ, are defined at a high level of resolution. This space is characterized by two key parameters: the extension Δx of its field, and its resolution scale δx. To define this *object space* more explicitly, we first introduce the finite grid (see Fig. 14.2):

$$\boldsymbol{G} = \boldsymbol{L} \times \boldsymbol{L}, \quad \boldsymbol{L} = \left\{ p \in \boldsymbol{Z} : -\frac{N}{2} \leq p \leq \frac{N}{2} - 1 \right\}, \tag{14.1}$$

where N is some power of 2.

On each *pixel* $\mathbf{p}\,\delta x (\mathbf{p} \in \boldsymbol{G})$, we then center a *scaling function* of the form

$$e_{\mathbf{p}}(\mathbf{x}) = e_{\mathbf{0}}(\mathbf{x} - \mathbf{p}\,\delta x) \quad \text{with} \quad e_{\mathbf{0}}(\mathbf{x}) = \operatorname{sinc}(\frac{x}{\delta x})\operatorname{sinc}(\frac{y}{\delta x}). \tag{14.2}$$

It is easy to verify that these functions form an orthogonal set. In this presentation of WIPE, the *object space* H_o is the Euclidian space generated by the basis vectors $e_{\mathbf{p}}$, $\mathbf{p}$ spanning $\boldsymbol{G}$ (see Fig. 14.2). The

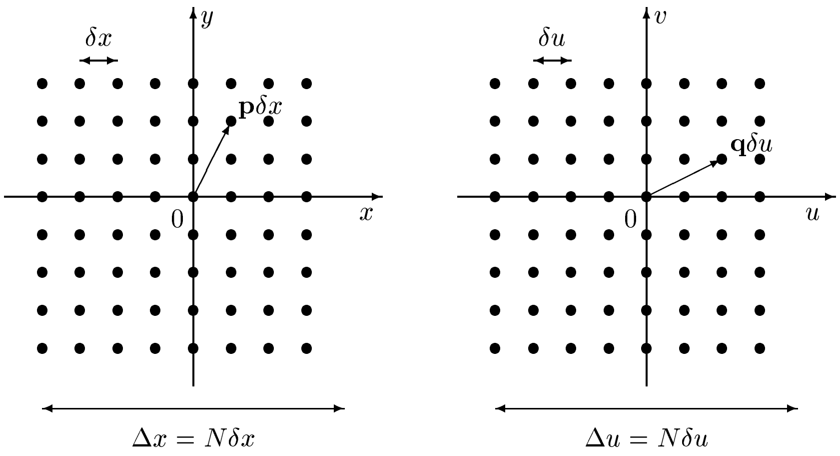


Figure 14.2: Object grid $\boldsymbol{G}\,\delta x$ *(left hand)* and Fourier grid $\boldsymbol{G}\,\delta u$ *(right hand)* for $N = 8$. The object domain is characterized by its resolution scale δx and the extension of its field $\Delta x = N\delta x$, where N is some power of 2 (the larger is N, the more oversampled is the object field). The basic Fourier sampling interval is $\delta u = 1/\Delta x$, the extension of the Fourier domain is $\Delta u = 1/\delta x$.

dimension of this space is equal to N^2: the number of *pixels* in the grid $\boldsymbol{G}$. The functions ϕ lying in H_o can therefore be expanded in the form

$$\phi(\mathbf{x}) = \sum_{\mathbf{p}\in\boldsymbol{G}} a_{\mathbf{p}} e_{\mathbf{p}}(\mathbf{x}), \tag{14.3}$$

where the $a_{\mathbf{p}}$'s are the components of ϕ in the interpolation basis of H_o.

The Fourier transform of ϕ is defined by the relationship

$$\widehat{\phi}(\mathbf{u}) = \int \phi(\mathbf{x})\,\mathrm{e}^{-2\mathrm{i}\pi\mathbf{u}\cdot\mathbf{x}}\,\mathrm{d}\mathbf{x},$$

where $\mathbf{u}$ is a two-dimensional angular spatial frequency: $\mathbf{u} = (u, v)$. According to the expansion of ϕ we therefore have:

$$\widehat{\phi}(\mathbf{u}) = \sum_{\mathbf{p}\in\boldsymbol{G}} a_{\mathbf{p}} \widehat{e}_{\mathbf{p}}(\mathbf{u}), \tag{14.4}$$

where

$$\widehat{e}_{\mathbf{p}}(\mathbf{u}) = \widehat{e}_{\mathbf{0}}(\mathbf{u})\,\mathrm{e}^{-2\mathrm{i}\pi\mathbf{p}\cdot\dfrac{\mathbf{u}}{\Delta u}} \quad \text{with} \quad \widehat{e}_{\mathbf{0}}(\mathbf{u}) = \frac{1}{(\Delta u)^2}\,\mathrm{rect}\Big(\frac{u}{\Delta u}\Big)\mathrm{rect}\Big(\frac{v}{\Delta u}\Big) \tag{14.5}$$

and $\Delta u = 1/\delta x$.

The dual space of the *object space*, $\widehat{H}_o$, is the image of H_o by the Fourier transform operator: $\widehat{H}_o$ is the space of the Fourier transforms of the functions ϕ lying in H_o. This space is characterized by two key parameters: its extension $\Delta u = 1/\delta x$, and the basic Fourier sampling interval $\delta u = 1/\Delta x$ (see Fig. 14.2).

14.3 Experimental data space

The *experimental data* $\Psi_e(\mathbf{u})$ are blurred values of $\widehat{\Phi}_o(\mathbf{u})$ on a finite list of frequencies in the Fourier domain:

$$\mathcal{L}_e = \{\mathbf{u}_e(1), \mathbf{u}_e(2), \ldots, \mathbf{u}_e(N_e)\}. \tag{14.6}$$

As the *object function* of interest Φ_o is a real-valued function, it is natural to define $\Psi_e(-\mathbf{u})$ as the complex conjugate of $\Psi_e(\mathbf{u})$. The *experimental frequency list* $\mathcal{L}_e$ is defined consequently: if $\mathbf{u} \in \mathcal{L}_e$, then $-\mathbf{u} \in \mathcal{L}_e$ (except for the null frequency $\mathbf{u} = \mathbf{0}$: in the convention adopted here, either it does not lie in $\mathcal{L}_e$, or there exists only one occurrence of this point). The *experimental frequency coverage* generated by $\mathcal{L}_e$ is therefore centrosymmetric (see Fig. 14.3).

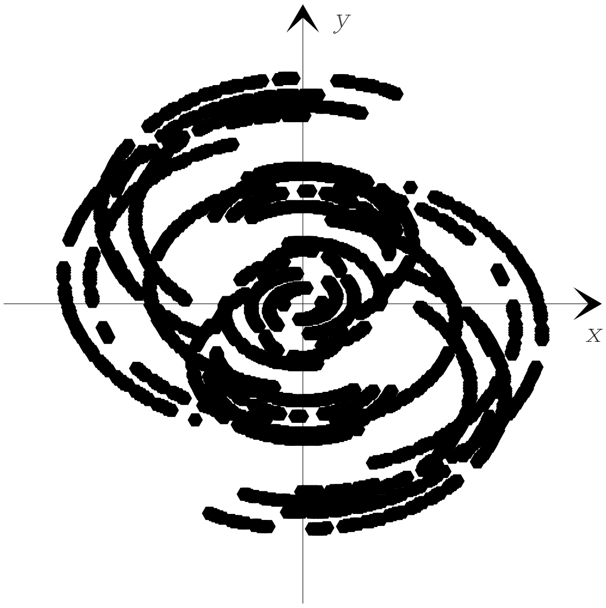


Figure 14.3: An example of an *experimental frequency coverage* provided by the IRAM interferometer. Here, the number of points N_e in the *experimental frequency list* $\mathcal{L}_e$ is equal to 2862.

The *experimental data* vector Ψ_e lies in the *experimental data space* K_e, the real Euclidian space underlying the space of complex-valued functions ψ on $\mathcal{L}_e$, such that $\psi(-\mathbf{u}) = \bar{\psi}(\mathbf{u})$. The dimension of this space is equal to N_e: the number of points in the *experimental frequency list* $\mathcal{L}_e$.

14.4 Image reconstruction process

As the *experimental frequency list* is finite, and in addition the *experimental data* blurred, the object representation that can be obtained from these data is of course incomplete. This simple remark shows that the inverse problems of Fourier synthesis must be regularized: the high-frequency components of the *image to be reconstructed* must be negligible.

The central problem is to specify in which conditions it is possible to extrapolate or interpolate, in some region of the Fourier domain, the Fourier transform of a function ϕ whose support is contained in some finite region of H_o. It is now well established that extrapolation is forbidden, and interpolation allowed to a certain extent. The corresponding regularization principle is then intimately related to the concept of resolution: the interpolation is performed in the frequency gaps of the *frequency coverage to be synthesized.*

14.4.1 Synthesized aperture

Let $\mathcal{H}$ be the Fourier domain: $\mathcal{H} = (-\Delta u/2, \Delta u/2)^2$. In Fourier synthesis, the *frequency coverage to be synthesized* is a centro-symmetric region $\mathcal{H}_s \subset \mathcal{H}$ (see Fig. 14.4).

CLEAN and WIPE share a common objective, that of the *image to be reconstructed.* This image, Φ_s, is defined so that its Fourier transform is quadratically negligible outside $\mathcal{H}_s$. More explicitly, Φ_s is defined by the convolution relation:

$$\Phi_s = \Theta_s \star \Phi_o. \tag{14.7}$$

The "synthetic beam" Θ_s is a function resulting from the choice of $\mathcal{H}_s$: the well-known *clean beam* in CLEAN, the *neat beam* in WIPE.

14.4.2 Synthetic beam

The *neat beam* can be regarded as a sort of optimal *clean beam*: the optimal apodized point-spread function that can be designed within the limits of the Heisenberg principle. More precisely, the *neat beam* Θ_s is a centro-symmetric function lying in the *object space* H_o, and satisfying the following properties:

- The energy of $\widehat{\Theta}_s$ is concentrated in $\mathcal{H}_s$. In other words, $\widehat{\Theta}_s$ has to be small outside $\mathcal{H}_s$ in the mean-square sense: we impose the fraction χ^2 of this energy in $\mathcal{H}_s$ to be close to 1 (say $\chi^2 = 0.98$).

- The effective support $\mathcal{D}_s$ of Θ_s in H_o is as small as possible with respect to the choice of $\mathcal{H}_s$ and χ^2. The idea is of course to have the best possible resolution.

This apodized point-spread function is thus computed on the grounds of a trade-off between resolution and efficiency, with the aid of the *power method*.

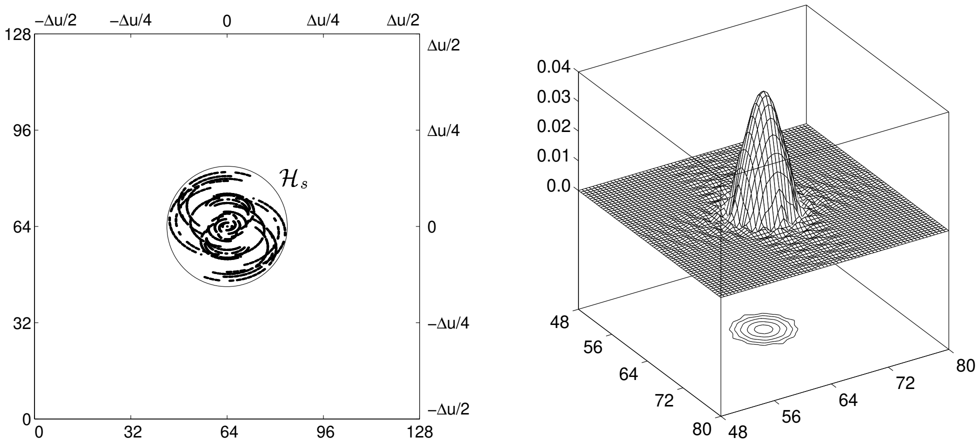


Figure 14.4: *Experimental frequency coverage* and *frequency coverage to be synthesized* $\mathcal{H}_s$ *(left hand)*. The *experimental frequency list* $\mathcal{L}_e$ includes $N_e = 2862$ frequency points. The *frequency coverage to be synthesized* $\mathcal{H}_s$ is centred in the Fourier grid $\boldsymbol{G}\,\delta u$, where $\delta u = \Delta u/N$ with $N = 128$ (here, the diameter of the circle is equal to $40\,\delta u$). The *neat beam* Θ_s *(right hand)* represented here corresponds to the *frequency coverage to be synthesized* $\mathcal{H}_s$ for a given value of $\chi^2 = 0.97$. It is centred in the object grid $\boldsymbol{G}\,\delta x$ where $\delta x = 1/\Delta u$ (here, the full width of Θ_s at half maximum is equal to $5\,\delta x$).

14.4.3 Regularization frequency list

As extrapolation is forbidden, and interpolation only allowed to a certain extent in the frequency gaps of the *frequency coverage to be synthesized*, the *experimental frequency list* $\mathcal{L}_e$ should be completed by high-frequency points. These points, located outside the *frequency coverage to be synthesized* $\mathcal{H}_s$, are those for which the high-frequency components of the *image to be reconstructed* are practically negligible.

The elements of the *regularization frequency list* $\mathcal{L}_r$ are the frequency points $\mathbf{u}_r$ located outside the *frequency coverage to be synthesized* $\mathcal{H}_s$ at the nodes of the Fourier grid $\boldsymbol{G}\,\delta u$:

$$\mathcal{L}_r = \{\mathbf{u}_r = \mathbf{q}\,\delta u, \mathbf{q} \in \boldsymbol{G} : \mathbf{q}\,\delta u \notin \mathcal{H}_s\}. \tag{14.8}$$

The *global frequency list* $\mathcal{L}$ is then the concatenation of $\mathcal{L}_e$ with $\mathcal{L}_r$.

14.4.4 Data space

According to the definition of the *image to be reconstructed*, the Fourier data corresponding to Φ_s are defined by the relationship:

$$\Psi_s(\mathbf{u}) = \widehat{\Theta}_s(\mathbf{u})\Psi_e(\mathbf{u}) \quad \forall \mathbf{u} \in \mathcal{L}_e. \tag{14.9}$$

Clearly, Ψ_s lies in the *experimental data space* K_e.

Let us now introduce the *data vector*:

$$\Psi_d(\mathbf{u}) = \begin{cases} \Psi_s(\mathbf{u}) & \text{on } \mathcal{L}_e; \\ 0 & \text{on } \mathcal{L}_r. \end{cases} \tag{14.10}$$

This vector lies in the *data space* K_d, the real Euclidian space underlying the space of complex-valued functions ψ on $\mathcal{L}$, such that $\psi(-\mathbf{u}) = \bar{\psi}(\mathbf{u})$. This space is equipped with the scalar product:

$$(\psi_1 \mid \psi_2)_d = \sum_{\mathbf{u}\in\mathcal{L}_e} \bar{\psi_1}(\mathbf{u})\psi_2(\mathbf{u})W(\mathbf{u})(\delta u)^2 + \sum_{\mathbf{u}\in\mathcal{L}_r} \bar{\psi_1}(\mathbf{u})\psi_2(\mathbf{u})(\delta u)^2; \tag{14.11}$$

$W(\mathbf{u})$ is a given *weighting function* that takes into account the reliability of the data via the standard deviation $\sigma_e(\mathbf{u})$ of $\Psi_e(\mathbf{u})$, as well as the local redundancy $\rho(\mathbf{u})$ of $\mathbf{u}$ up to the sampling interval δu.

The *Fourier sampling operator* A is the operator from the *object space* H_o into the *data space* K_d:

$$A : H_o \longrightarrow K_d, \qquad (A\phi)(\mathbf{u}) = \begin{cases} \widehat{\phi}(\mathbf{u}) & \text{on } \mathcal{L}_e; \\ \widehat{\phi}(\mathbf{u}) & \text{on } \mathcal{L}_r. \end{cases} \tag{14.12}$$

As the *experimental data* $\Psi_e(\mathbf{u})$ are blurred values of $\widehat{\Phi}_o(\mathbf{u})$ on $\mathcal{L}_e$, this operator will play a key role in the image reconstruction process. The definition of this *Fourier sampling operator* suggests that the action of A should be decomposed into two components: A_e on the *experimental frequency list* $\mathcal{L}_e$, and A_r on the *regularization frequency list* $\mathcal{L}_r$.

14.4.5 Object representation space

The *reconstructed image* is defined as the function Φ_E of the *object space* H_o minimizing some objective functional. The definition of this functional takes into account the nature of the data, as well as other constraints. For example, the *image to be reconstructed* may be confined to a subspace, or more generally to a convex set, of the *object space* H_o: this convex set is the *object representation space* E. It may be defined from the outset (in an interactive manner, for example), or step by step throughout the image reconstruction procedure (this is the case of the current implementation of WIPE). In both cases, the projection operator onto this space, the projector P_E, will play an essential role in the image reconstruction process.

REMARK 1: positivity constraint.

In most cases encountered in practice, the scalar components of Φ_E in the interpolation basis of H_o must be non-negative (cf. Eq.14.2). In the current implementation of WIPE this constraint is taken into account. The *object representation space* E is then built, step by step, accordingly.

14.4.6 Objective functional

The *reconstructed image* is defined as the function Φ_E minimizing on E the objective functional:

$$q(\phi) = \|\Psi_d - A\phi\|_d^2. \tag{14.13}$$

According to the definition of the *data vector* Ψ_d and to that of the *Fourier sampling operator* A, this quantity can be written in the form:

$$q(\phi) = q_e(\phi) + q_r(\phi) \text{ with } \begin{cases} q_e(\phi) = \displaystyle\sum_{\mathbf{u}\in\mathcal{L}_e} |\Psi_s(\mathbf{u}) - \widehat{\phi}(\mathbf{u})|^2 W(\mathbf{u})(\delta u)^2; \\ q_r(\phi) = \displaystyle\sum_{\mathbf{u}\in\mathcal{L}_r} |\widehat{\phi}(\mathbf{u})|^2 (\delta u)^2. \end{cases} \tag{14.14}$$

The experimental criterion q_e constraints the *object model* ϕ to be consistent with the damped Fourier data Ψ_s, while the regularization criterion q_r penalizes the high-frequency components of ϕ.

Let now F be the image of E by A (the space of the $A\phi$'s, ϕ spanning E), A_E be the operator from E into F induced by A, and Ψ_F the projection of Ψ_d onto F (see Fig. 14.7). The vectors ϕ minimizing q on E, the solutions of the problem, are such that $A_E\phi = \Psi_F$. They are identical up to a vector lying in the kernel of A_E (by definition, the kernel of A_E is the space of vectors ϕ such that $A_E\phi = 0$).

As $\Psi_d - \Psi_F$ is orthogonal to F, the solutions ϕ of the problem are characterized by the property: $\forall\varphi \in E, (A\varphi \mid \Psi_d - A\phi)_d = 0$. On denoting by A^* the adjoint of A, this property can also be written in the form:

$$\forall\varphi \in E, \quad (\varphi \mid r)_o = 0, \text{ with } r = A^*(\Psi_d - A\phi). \tag{14.15}$$

where r is regarded as a residue. This condition is of course equivalent to $P_E r = 0$, where P_E is the projector onto the *object representation space* E. The solutions of the problem are therefore the solutions of the *normal equation* on E:

$$A_E^* A_E \phi = A_E^* \Psi_d, \tag{14.16}$$

where $A_E^* = P_E A^*$.

Many different techniques can be used for solving the *normal equation* (or minimizing q on E). Some of these are certainly more efficient than others, but this is not a crucial choice.

REMARK 2: beams and maps.

The action of A^*A involved in $A_E^* A_E$ is that of a convolutor. As the two lists $\mathcal{L}_e$ and $\mathcal{L}_r$ are disjoints, we have: $A^*A = A_e^* A_e + A_r^* A_r$. Thus, the corresponding point-spread function, called the *dusty beam*, has two components: the traditional *dirty beam* Θ_d and the *regularization beam*. The latter corresponds to the action of $A_r^* A_r$, the former to that of $A_e^* A_e$ (see Fig. 14.5). Likewise, according to the definition of the *data vector*, $A^*\Psi_d = A_e^*\Psi_s$ is called the *dusty map* (as opposed to the traditional *dirty map* $A_e^*\Psi_d$ because it is damped by the *neat beam*).

REMARK 3: construction of the *object representation space*.

With regard to the construction of the *object representation space* E, CLEAN and WIPE are very similar: it is defined through the choice of the (discrete) object support. It is important to note that this space may be constructed, in a global manner or step by step, interactively or automaticaly. In the last version of WIPE implemented at IRAM, the image reconstruction process is initialized with a few iterations of CLEAN. The support selected by CLEAN is refined throughout the iterations of WIPE by conducting a matching pursuit process at the level of the components of r in the interpolation basis of H_o: the current support is extended by adding the nodes of the object grid $\boldsymbol{G}\,\delta x$ for which these coefficients are the largest above a given threshold (half of the maximum value, for example). The objective functional is then minimized on that new support, and the global residue r updated accordingly. The *object representation space* of the *reconstructed image* is thus obtained step by step in a natural manner.

The simulation presented on Fig.14.5-14.6 corresponds to the conditions of Fig. 14.4. The Fourier data Ψ_e were blurred by adding a Gaussian noise: for all $\mathbf{u} \in \mathcal{L}_e$, the standard deviation of $\Psi_e(\mathbf{u})$ was set equal to 5% of the total flux of the object $(\widehat{\Phi}_o(\mathbf{0})/20)$. The image reconstruction process was initialized with a few iterations of CLEAN, and the construction of the final support of the *reconstructed image* was made as indicated in Remark 3. At the end of the reconstruction process, a final smoothing of the current object support was performed. In this classical operation of mathematical morphology, the effective support of Θ_s, $\mathcal{D}_s$, is of course used as a structuring element. The boundaries of the effective support of the reconstructed *neat map* are thus defined at the appropriate resolution. In particular, the connected entities of size smaller than that of $\mathcal{D}_s$ are eliminated.

14.4.7 Uniqueness and robustness

When the problem is well-posed, A_E is a one-to-one map ($\ker A_E = \{0\}$) from E onto F; the solution is then unique: there exists only one vector $\phi \in E$ such that $A_E\phi = \Psi_F$. This vector, Φ_E, is said to be the least-squares solution of the equation $A_E\phi$ "=" Ψ_d.

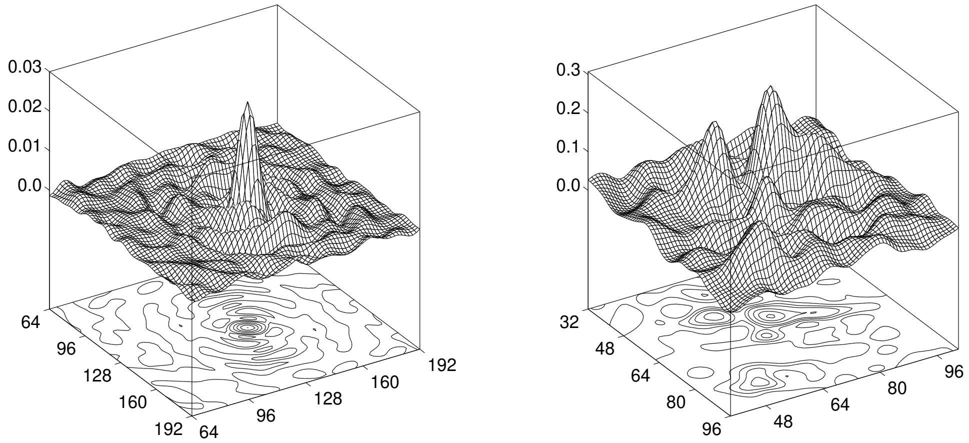


Figure 14.5: *Dirty beam (left hand)* corresponding to the *experimental frequency list* $\mathcal{L}_e$ of Fig. 14.4, and *dusty map (right hand)* of a simulated data set (the simulated Fourier data Ψ_e were blurred by adding a Gaussian noise with a standard deviation σ_e equal to 5% of the total flux of the object Φ_o).

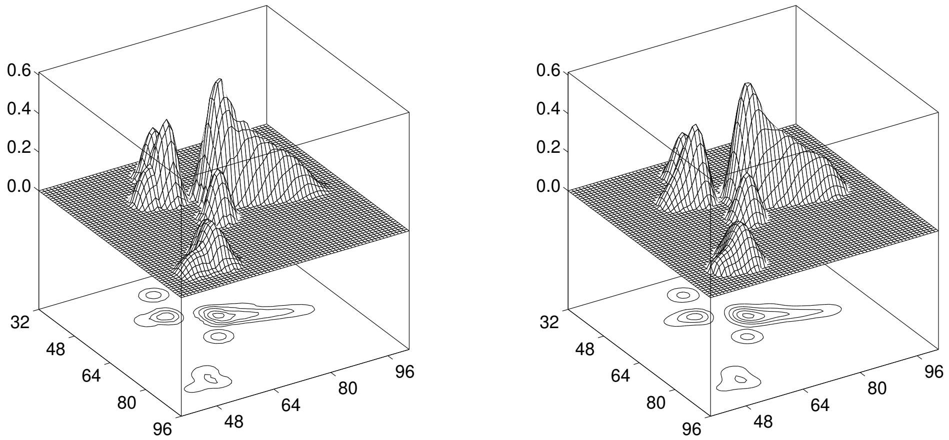


Figure 14.6: *Image to be reconstructed* Φ_s *(left hand)* at the resolution level defined in Fig. 14.4, and reconstructed *neat map* Φ_E *(right hand)* at the same resolution: the final *condition number* κ_E is equal to 2.46 (cf. Eq. 14.17 and 14.18).

In this case, let $\delta\Psi_F$ be a variation of Ψ_F in F, and $\delta\Phi_E$ be the corresponding variation of Φ_E in E (see Fig. 14.7). It is easy to show that the robustness of the reconstruction process is governed by the inequality:

$$\frac{\|\delta\Phi_E\|_o}{\|\Phi_E\|_o} \leq \kappa_E \frac{\|\delta\Psi_F\|_d}{\|\Psi_F\|_d}. \tag{14.17}$$

The error amplifier factor κ_E is the condition number of A_E:

$$\kappa_E = \frac{\sqrt{\lambda'}}{\sqrt{\lambda}}; \tag{14.18}$$

here λ and λ' respectively denote the smallest and the largest eigenvalues of $A_E^* A_E$. The closer to 1 is the condition number, the easier and the more robust is the reconstruction process (see Fig. 14.8 and 14.9).

The part played by inequality 14.17 in the development of the corresponding error analysis shows that a good reconstruction procedure must also provide, in particular, the condition number κ_E. This is the

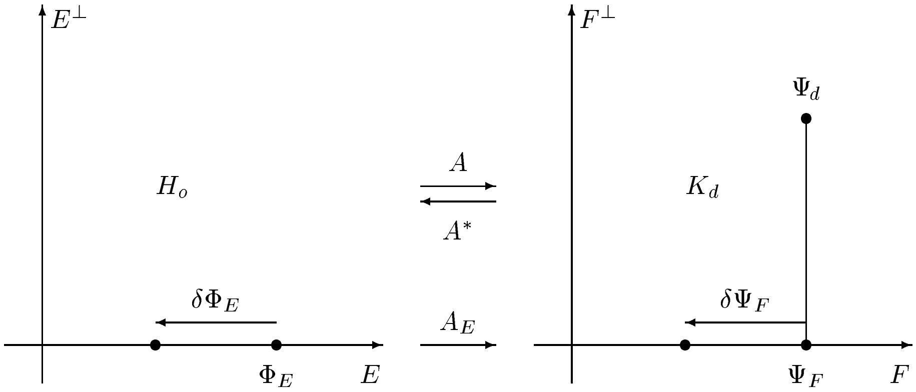


Figure 14.7: Uniqueness of the solution and robustness of the reconstruction process. Operator A is an operator from the *object space* H_o into the *data space* K_d. The *object representation space* E is a particular subspace of H_o. The image of E by A, the range of A_E, is denoted by F. In this representation, Ψ_F is the projection of the data vector Ψ_d onto F. The inverse problem must be stated so that A_E is a one-to-one map from E onto F, the condition number κ_E having a reasonable value.

case of the current implementation of WIPE which uses the *conjugate gradient* method for solving the *normal equation* 14.16.

To conduct the final error analysis, one is led to consider the eigenvalue decomposition of $A_E^* A_E$. This is done, once again, with the aid of the *conjugate gradient* method associated with the *QR* algorithm. At the cost of some memory overhead (that of the M successive residues), the latter also yields approximations of the eigenvalues λ_k of $A_E^* A_E$. It is thus possible to obtain the scalar components of the associated eigenmodes Φ_k in the interpolation basis of H_o. The purpose of this analysis is to check whether some of them (in particular those corresponding to the smallest eigenvalues) are excited or not in Φ_E. If so, the corresponding details may be artefacts of the reconstruction.

The reconstructed map is then decomposed in the form:

$$\Phi_E = \sum_{k=1}^{M} w_k \Phi_k, \quad w_k = (\Phi_k \mid \Phi_E). \tag{14.19}$$

The separation angle θ_k between Φ_E and Φ_k is explicitly given by the relationship:

$$\cos\theta_k = \frac{w_k}{\sqrt{\sum_{k=1}^{M} w_k^2}} \quad (0 \leq \theta_k \leq \pi/2). \tag{14.20}$$

The closer to $\pi/2$ is θ_k, the less excited is the corresponding eigenmode Φ_k in the reconstructed *neat map* Φ_E.

To illustrate in a concrete manner the interest of equations 14.19 and 14.20, let us consider the simulations presented in Fig. 14.4 and 14.9. Whatever the value of the final condition number is, the error analysis allows the astronomer to check if there exists a certain similitude between some details in the *neat map* and some features of the critical eigenmodes. This information is very attractive, in particular when the resolution of the reconstruction process is greater than a reasonable value (the larger is the *aperture to be synthesized* $\mathcal{H}_s$, the smaller is the full width at half-maximum of Θ_s). In such situations of "super resolution," the error analysis will suggest the astronomer to redefined the problem at a lower level of resolution, or to keep in mind that some details in the reconstructed *neat map* may be artefacts of the reconstruction process.

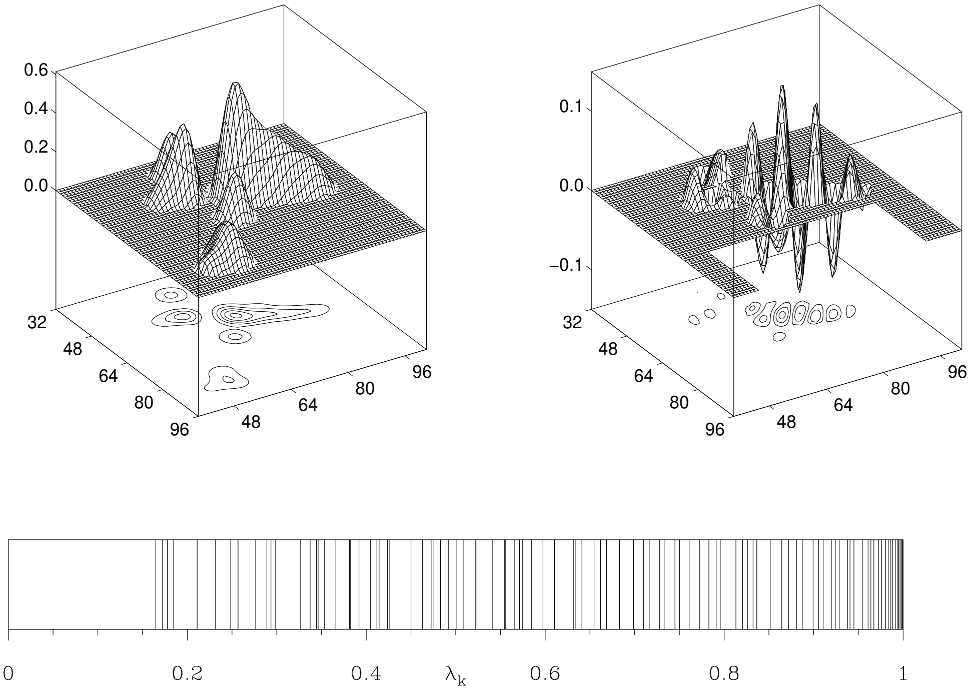


Figure 14.8: Reconstructed *neat map* Φ_E *(left hand)* and eigenmode Φ_1 *(right hand)* corresponding to the smallest eigenvalue $\lambda_1 = 0.165$ of $A_E^* A_E$. The conditions of the simulations are those of Fig. 14.4 and 14.5: in particular, the diameter of $\mathcal{H}_s$ is equal to $40\,\delta u$. The final condition number is $\kappa_E = 2.46$ (the eigenvalues of $A_E^* A_E$ are plotted on the bar code below). This eigenmode is not excited in Φ_E: the separation angle θ_1 between Φ_E and Φ_1 is greater than 89°. In other situations, when the final condition number is greater, this mode may be at the origin of some artefacts in the *neat map* (see Fig. 14.9).

14.5 Implementation of WIPE at IRAM

In this section we describe the successive steps of the image reconstruction process as it is implemented now in the MAPPING program included in the IRAM software. For more information on this program, the reader is invited to read the last version of the *Mapping CookBook*.

The first step of the image reconstruction process is to defined the *object space* H_o. This space is characterized by two key parameters: the extension Δx of its field, and its resolution scale $\delta x = \Delta x / N$ (see Fig. 14.2). The procedure `wipe_init` is used to set these parameters properly.

The *frequency coverage to be synthesized* $\mathcal{H}_s$ is defined with the aid of the procedure `wipe_aper`. This tool provides an interactive way of fitting an ellipse over the *experimental frequency coverage* generated by the *experimental frequency list* $\mathcal{L}_e$ (see Fig. 14.4).

Once $\mathcal{H}_s$ has been defined, the procedure `wipe_beam` is ready for computing the *neat beam* Θ_s, as well as the *dirty beam* Θ_d. The latter plays a key role in the action of the convolutor $A_E^* A_E$, while the Fourier transform of the former is involved in the definition of the *data vector* Ψ_d (cf. Eq. 14.9 and 14.10).

The last step in the image reconstruction process concerns the *neat map*. It is implemented in the `wipe_solve` command. Before the initialization of the reconstruction, the *dusty map* $A^* \Psi_d$ is computed, and an optional support can be selected (this support plays the role of the *clean box* of CLEAN). As WIPE can be slow when reconstructing large images, it can be initialized with a few CLEAN iterations to quickly build a first *object representation space* E. When switching to WIPE, the program starts by optimizing the solution provided by CLEAN with the corresponding support. Then, at each iteration of WIPE, the support grows, and for a given and fixed *object representation space* E, the *normal equation* 14.16 is solved by using the *conjugate gradient* method, which also provides the condition number κ_E of A_E. When leaving WIPE, a final smoothing of the current object support is performed, removing (through an appropriate morphological analysis) the details of the *reconstructed image* smaller than the resolution limit

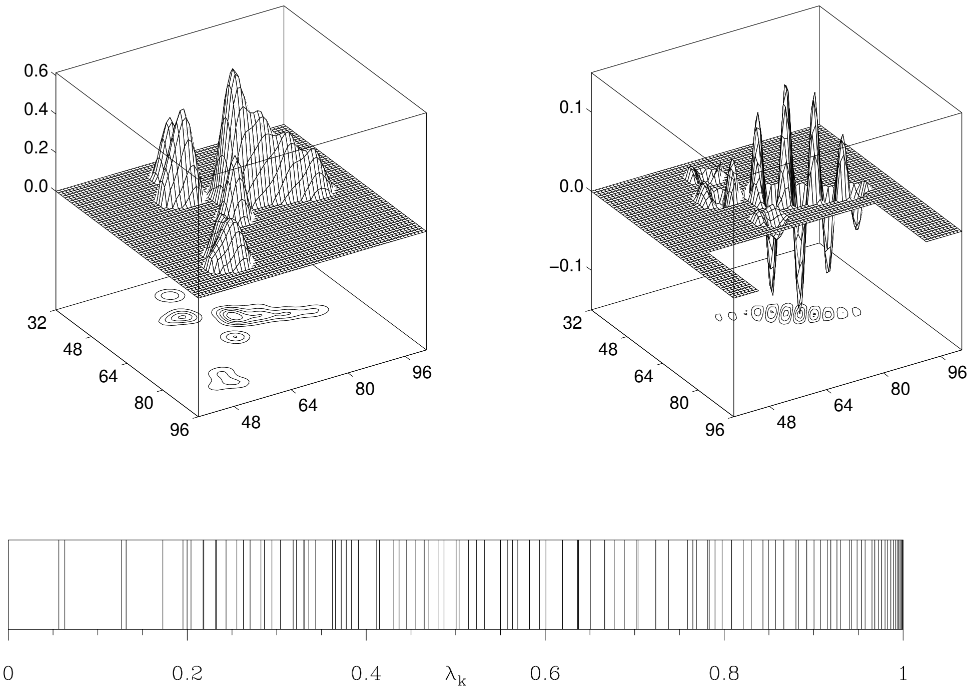


Figure 14.9: Reconstructed *neat map* Φ_E *(left hand)* and related critical eigenmode Φ_1 *(right hand)*. The latter corresponds to the smallest eigenvalue $\lambda_1 = 0.057$ of $A_E^* A_E$. The conditions of the simulations are those of Fig. 14.5, but here the diameter of $\mathcal{H}_s$ is taken equal to $48\,\delta u$: the final condition number is $\kappa_E = 4.19$ (the eigenvalues of $A_E^* A_E$ are plotted on the bar code below). The critical eigenmode Φ_1 is at the origin of the oscillations along the main structuring entity of Φ_E. This mode is slightly excited (the separation angle θ_1 between Φ_E and Φ_1 is less than 86°), thus the corresponding details may be artefacts. In this case of "super-resolution" the error analysis provided by WIPE suggests that the procedure should be restarted at a lower level of resolution (see Fig. 14.8), so that the final solution be more stable and reliable.

of the reconstruction process. The final *reconstructed image* Φ_E is the function minimizing the objective functional 14.13 on that support.

The control of the robustness of the reconstruction process is performed through an additional step with the `wipe_error` command. This procedure computes with a fine accuracy the final *condition number* κ_E, as well as the eigenvalues and the critical eigenmodes of $A_E^* A_E$. One of the aims of this last step is to check that the features present in the *reconstructed image* are not artefacts. This can be done by comparing these features with those of the critical eigenmodes. When there exists a certain similitude (between these features), it is then recommended to restart the process with a lower resolution, so that the final solution be more stable and reliable.

Glossary

$\boldsymbol{L}$, N	One-dimensional grid, number of elements in $\boldsymbol{L}$
$\boldsymbol{G} = \boldsymbol{L} \times \boldsymbol{L}$	Two-dimensional grid
$\mathbf{p} = (p, q)$	Two-dimensional integer vector
$\mathbf{x} = (x, y)$	Two-dimensional angular position variable
$\mathbf{u} = (u, v)$	Two-dimensional angular spatial frequency
Δx	Extension of the synthesized field
δx	Resolution scale of the synthesized field
Δu	Extension of the Fourier domain
δu	Basic Fourier sampling interval
$\boldsymbol{G}\,\delta x$, $\boldsymbol{G}\,\delta u$	Object grid, Fourier grid
$\mathcal{L}$	Global frequency list
$\mathcal{L}_e$, $\mathcal{L}_r$	Experimental frequency list, regularization frequency list
$\mathcal{H}$	Fourier domain $[-\Delta u/2, \Delta u/2]$
$\mathcal{H}_s$	Frequency coverage to be synthesized
$\mathcal{D}_s$	Support of the neat beam Θ_s
χ^2	Energy confinement parameter
Θ_s	Apodized point-spread function (neat beam)
Θ_d	Instrumental point-spread function (dirty beam)
H_o, $e_{\mathbf{p}}(\mathbf{x})$	Object space, basis functions of H_o
E, F	Object representation space, image of E by A
K_e, K_d	Experimental data space, data space
$W(\mathbf{u})$	Weighting function
$\rho(\mathbf{u})$, $\sigma_e(\mathbf{u})$	Redundancy of $\mathbf{u}$, standard deviation of $\Psi_e(\mathbf{u})$
A	Regularized Fourier sampling operator
A_e, A_r	Fourier sampling operator on $\mathcal{L}_e$, on $\mathcal{L}_r$
P_E, A_E	Projection operator onto E, restriction of A to E
Φ_o, Φ_s	Original object function, image to be reconstructed
Φ_E, $\delta\Phi_E$	Reconstructed image, reconstruction error on Φ_E
Ψ_e, Ψ_s	Experimental data, damped experimental data
Ψ_d	Regularized data vector
Ψ_F, $\delta\Psi_F$	Projection of Ψ_d onto F, effective error on Ψ_F
λ_k, Φ_k	Eigenvalue of $A_E^* A_E$ and related eigenmode
θ_k	Separation angle between Φ_E and Φ_k
λ, λ'	Smallest and largest eigenvalues of $A_E^* A_E$
κ_E	Condition number of A_E
$q(\phi)$	Regularized criterion
$q_e(\phi)$, $q_r(\phi)$	Experimental criterion, regularization criterion

Chapter 15

Mosaicing

Fréderic Gueth

Max Planck Institute für Radioastronomie, Auf dem Hügel 69, D-5300 Bonn

15.1 Introduction

The dirty map F resulting from a normal, single-field interferometric observation can be described by the equation:

$$F = D * (B \times I) + N \tag{15.1}$$

where D is the dirty beam, B the antenna primary beam, I the sky brightness distribution, and N a noise distribution[1]. The interferometer is only sensitive to the product of the sky brightness distribution by the rapidly decreasing function B. As the noise distribution N is not affected by B, any attempt to correct for the primary beam attenuation (i.e. divide the clean map by B) results in a strongly increasing noise. Hence, the primary beam attenuation limits the size of the region it is possible to map with an interferometer.

Due to the coupling between the receiver horn and the primary mirror of the antennas (see lecture by A. Greve), the primary beam B is, to a good approximation, a gaussian. Its FWHM (proportional to the ratio of the wavelength λ to the antenna diameter $\mathcal{D}$) can therefore be used to define a "field of view". Table 15.1 gives the resulting values for the Plateau de Bure interferometer, for different frequencies. To map regions more extended than the primary beam width, it is necessary to observe a *mosaic* of several adjacent fields. Clearly, due to the gaussian-shape of the primary beam attenuation, these fields have to strongly overlap to ensure a roughly uniform sensitivity over the whole mapped region.

A further complication arises from the lack of the short-spacings information in the interferometer data set. Due to their diameter, the antennas cannot be put too close to each other, which results in a minimal measured baseline (24 m at the Plateau de Bure). Even if projection effects reduce the effective baselines, a central "hole" in the data distribution in the uv plane cannot be avoided. As a consequence, the extended structures (whose visibilities are confined in a small region in the uv plane) are filtered

[1] In the following, we will assume an uniform noise rms, i.e. we do not take into account variation of the noise introduced by the imaging process (see lecture by S. Guilloteau).

Frequency (GHz)	Wavelength (mm)	Field of View ($''$)	Largest structure ($''$)
85	3.5	58	36
100	3.0	50	31
115	2.6	43	27
215	1.4	23	14
230	1.3	21.5	13
245	1.2	20	12

Table 15.1: Field of view of the Plateau de Bure interferometer: the 15 m dishes have a gaussian illumination, which yields a nearly gaussian primary beam. The two groups of frequencies correspond to the two receivers that are currently available. The last column gives *rough estimates* of the size of the largest structure which can be observed.

out. The largest structure it is possible to map with a single-field interferometric observation is thus even smaller than the field of view, and can be very roughly estimated by the ratio of the wavelength to the minimal baseline (Table 15.1). In the framework of mosaic observations, the short-spacings problem has however to be thought in slightly different terms, because it introduces now artifacts on an intermediate scale (see Sec. 15.5) but also because it can, at least in theory, be partially solved (Sec. 15.2).

15.2 Image formation in a mosaic

Some important mosaic properties can be understood by analyzing the combination of the data directly in the uv plane. This analysis was first proposed by [Ekers & Rots 1979]. The reader is also referred to [Cornwell 1989]. We consider a source whose brightness distribution is $I(x, y)$ (where x and m are two angular coordinates), and whose "true" visibility (i.e. the Fourier transform of I) is noted V. A two-antenna interferometer, whose primary beam is $B(x, y)$, will measure a visibility at a point (u, v) which may be written as:

$$V_{\mathrm{mes}}(u, v) = \iint_{-\infty}^{+\infty} B(x, y)\, I(x, y)\, \mathrm{e}^{-2i\pi(ux+vy)}\, dx\, dy \tag{15.2}$$

For an observation with a phase center in $(x = 0, m = 0)$ but with a pointing center in (x_p, y_p), the measured visibility (whose dependence on (x_p, y_p) is here explicitly indicated) is now:

$$V_{\mathrm{mes}}(u, v, x_p, y_p) = \iint_{-\infty}^{+\infty} B(x - x_p, y - y_p)\, I(x, y)\, \mathrm{e}^{-2i\pi(ux+vy)}\, dx\, dy \tag{15.3}$$

For further use, this last relation can also be rewritten, using the symmetry properties of the primary beam B:

$$V_{\mathrm{mes}}(u, v, x_p, y_p) = B(x_p, y_p) * \mathcal{F}(u, v, x_p, y_p) \tag{15.4}$$

where $*$ denotes a convolution product and the function $\mathcal{F}$ is defined as:

$$\mathcal{F}(u, v, x_p, y_p) = I(x_p, y_p)\, \mathrm{e}^{-2i\pi(ux_p+vy_p)} \tag{15.5}$$

Now, imagine an ideal "on-the-fly" mosaic experiment: for a given, fixed (u, v) point, the pointing center is continuously modified, and the variation of the visibility with $(x_p,\, y_p)$ can thus be monitored. Then, the Fourier transform of V_{mes} with respect to (x_p, y_p) gives (from equation 15.4):

$$[\mathrm{FT_p}(V_{\mathrm{mes}})]\,(u_p, v_p) = T(u_p, v_p)\, V(u + u_p, v + v_p) \tag{15.6}$$

where:

- $\mathrm{FT_p}$ denotes the Fourier transform with respect to (x_p, y_p) and (u_p, v_p) are the conjugate variables to (x_p, y_p).
- $[\mathrm{FT_p}(V_\mathrm{mes})]\,(u_p, v_p)$ is the Fourier transform of the observations.
- $T(u_p, v_p)$ is the Fourier transform of the primary beam $B(x_p, y_p)$. T is thus the transfer function of each antenna. For a dish of diameter $\mathcal{D}$, $T(u_p, v_p) = 0$ if $\sqrt{u_p^2 + v_p^2} > \mathcal{D}/\lambda$.
- $V(u + u_p, v + v_p)$ is the Fourier transform of $\mathcal{F}(u, v, x_p, y_p)$ with respect to (x_p, y_p). Indeed, $\mathcal{F}$ is the product of the sky brightness distribution (whose Fourier transform is V) by a phase term, and its own Fourier transform is thus simply V taken at a shifted point.

For $\sqrt{u_p^2 + v_p^2} < \mathcal{D}/\lambda$, we can thus derive:

$$V(u + u_p, v + v_p) = \frac{[\mathrm{FT_p}(V_\mathrm{mes})]\,(u_p, v_p)}{T(u_p, v_p)} \tag{15.7}$$

This relation illustrates an important property of the experiment we have considered. The observations were performed at a given (u, v) point but with a varying pointing center. Equation 15.7 shows that is possible to derive from this data set the visibility $V(u + u_p, v + v_p)$ at all (u_p, v_p) which verify $(u_p^2 + v_p^2)^{1/2} < \mathcal{D}/\lambda$. In other terms, the measurements have been done at (u, v) but the redundancy of the observations allows to compute (through a Fourier transform and a division by the antenna transfer function) the source visibility at all the points of a disk of radius $\mathcal{D}/\lambda$, centered in (u, v).

Interpretation

In very pictorial terms, one can say that the adjacent pointing reinforce each other and thereby yield a estimate of the source visibility at unmeasured points. Note however that the resulting image quality is not going to be drastically increased: more information can be extracted from the data, but a much more extended region has now to be mapped[2]. The redundancy of the observations has only allowed to rearrange the information in the uv-plane. This is nevertheless extremely important, as e.g. it allows the estimate part of the missing short-spacings (see below).

How is it possible to recover unmeasured spacings in the uv-plane? It is actually obvious that two antennas of diameter $\mathcal{D}$, separated by a distance $\mathcal{B}$, are sensitive to all the baselines ranging from $\mathcal{B} - \mathcal{D}$ to $\mathcal{B} + \mathcal{D}$. The measured visibility is therefore an average of all these baselines: V_mes is actually the convolution of the "true" visibility by the transfer function of the antennas. This is shown by the Fourier transform of equation 15.2, which yields: $V_\mathrm{mes} = T * V$. Now, if the pointing center and the phase center differ, a phase gradient is introduced across the antenna apertures, which means that the transfer function is affected by a phase term. Indeed, the Fourier transform of equation 15.3 yields:

$$V_\mathrm{mes}(u, v) = \left[T(u, v)\,\mathrm{e}^{-2i\pi(ux_p + vy_p)}\right] * V(u, v) \tag{15.8}$$

Hence, the measured visibilities are (still) a linear combination of the "true" visibilities. Measurements performed in various directions (x_p, y_p) give many such linear combinations. One can thus expect to derive from this linear system the initial visibilities, in the baseline range from $\mathcal{B} - \mathcal{D}$ to $\mathcal{B} + \mathcal{D}$. Equation 15.7 just shows that a Fourier transform allows to do that.

Field spacing in a mosaic

In the above analysis, a continuous drift of the pointing center was considered. However, the same results can be reached in the case of a limited number of pointings, provided that classical sampling theorems are fulfilled. We want to compute the visibility in a finite domain, which extends up to $\pm\mathcal{D}/\lambda$ around the nominal center, and therefore the pointing centers have to be separated by $\lambda/2\mathcal{D}$ (see [Cornwell 1988]).

[2] We have considered observations of different directions, performed with the same uv-coverage. The analysis presented here shows that such an experiment is somehow equivalent to the observation of the whole source, but with a different, more complete uv-coverage.

In practice, the (gaussian) transfer function of the millimeter dishes drops so fast that one can use without consequences a slightly broader, more convenient sampling, equal to half the primary beam width (i.e. $1.2\,\lambda/2\mathcal{D}$).

Mosaics and short-spacings

As with any other measured point in the uv plane, it is possible to derive visibilities in a small region (a disk of diameter $\mathcal{D}/\lambda$) around the shortest measured baseline. This is the meaning of the statement that mosaics can recover part of the short-spacings information: a mosaic will include (u, v) points corresponding to the shortest baseline minus $\mathcal{D}/\lambda$.

In practice, however, things are more complex. First, we have to deal with noisy data. As a consequence, it is not possible to expect a gain of $\mathcal{D}/\lambda$: the transfer function T which is used in equation 15.7 is strongly decreasing, and therefore signal-to-noise ratio limits the gain in the uv plane to a smaller value, typically $\mathcal{D}/2\lambda$ ([Cornwell 1988]). This is still a very useful gain: for the Plateau de Bure interferometer, this corresponds to a distance in the uv plane of 7.5 m, while the shortest (unprojected) baseline is 24 m. Secondly, the analysis described above would be rather difficult to implement with real observations. Instead, one prefers to combine the observed fields to directly reconstruct the sky brightness distribution. The resulting image should include the information arising from the redundancy of the adjacent fields, among them part of the short-spacings. However, the complexity of the reconstruction and deconvolution algorithms that have to be used precludes any detailed mathematical analysis of the structures of the maps. For instance, the (unavoidable) deconvolution of the image can also be interpreted as an interpolation process in the uv plane (see [Schwarz 1978]) for the case of the CLEAN algorithm) and its effects can thus hardly be distinguish from the intrinsic determination of unmeasured visibilities that occur when mosaicing.

15.3 Mosaicing in practice

Observation and calibration

The observation of a mosaic with the Plateau de Bure interferometer and the calibration of the data do not present any specific difficulties. We just mention here a few practical remarks:

- As shown in the previous paragraph, the optimal spacing between adjacent fields is half the primary beam width. Larger separations can be used (e.g. to map larger field of view in the same amount of time) but the image reconstruction is not optimal in that case. Note that if the two receivers are used simultaneously, the field spacing has to be adapted to one of the frequencies, which results in an over- or undersampling for the other one.

- Even if this is not formally required by the reconstruction and deconvolution algorithm described in the following section, it seems quite important to ensure similar observing conditions for all the pointing centers. Ideally, one would want the same noise level in each field, so that the noise in the final image is uniform, and the same uv coverage, to avoid strong discrepancies (in terms of angular resolution and image artifacts) between the different parts of the mosaic. To handle these constraints in practice, the fields are observed in a loop, each one during a few minutes (similarly to snapshot observations of several sources): hence, atmospheric conditions and uv coverage are similar for all the fields.

- In most cases, a mosaic will not be observed during an amount of time significantly larger than normal projects. As the observing time is shared between the different pointing centers, the sensitivity of each individual field is thus smaller than what would have been achieved with normal single-field observations. Note however that the sensitivity is further increased in the mosaic, thanks to the strong overlap between the adjacent fields (see Fig. 15.1).

- The maximal number of fields it is possible to observe in a mosaic is limited by observational constraints. The fields are observed in a loop, one after the other, and to get a reasonable uv

coverage within one transit, only a limited number of fields can be observed. With the Plateau de Bure interferometer, the limit seems to be around 15 fields. Mosaicing even more fields would probably require some other approach (e.g. mosaic of several mosaics). Finally, a potential practical limitation is the disk and memory sizes of the computers, as mosaicing requires to handle very large images.

- The calibration of a mosaic is strictly identical with any other observation performed with the Plateau de Bure interferometer, as only the observations of the calibrators are used. At the end of the calibration process, a *uv* table and then a dirty map are computed for each pointing center.

Mosaic reconstruction

The point is now to reconstruct a mosaic from the observations of each field, in an optimal way in terms of signal-to-noise ratio. Just forget for the time being the effects of the convolution by the dirty beam. Each field i can thus be written: $F_i = B_i \times I + N_i$, where B_i is the primary beam of the interferometer, centered in a different direction for each observation i, and N_i is the corresponding noise distribution. In practice, the same phase center (i.e. the same coordinate system) is used for all the fields. We are thus in the classical framework of several observations of the same unknown quantity I, each one being affected by a weighting factor B_i. The best estimate of I, in the least-square sense, is thus given by:

$$\widetilde{I} = \frac{\sum_i \frac{B_i}{\sigma_i^2} F_i}{\sum_i \frac{B_i^2}{\sigma_i^2}} \tag{15.9}$$

where the sum includes all the observed fields and σ_i is the rms of the noise distribution N_i. (Note that in Eq.15.9 as well as in the following equations, σ_i is a number while other letters denote two-dimensional distributions).

Linear vs. non-linear mosaics

The problem which remains to be address is the deconvolution of the mosaic. This is actually the main difficulty of mosaic interferometric observations. Two different approaches have been proposed (e.g. [Cornwell 1993]):

- *Linear mosaicing:* each field is deconvolved using classical technics, and a mosaic is reconstructed afterwards with the clean images, according to equation 15.9.

- *Non-linear mosaicing:* a joint deconvolution of all the fields is performed, i.e. the reconstruction and the deconvolution of the mosaic are done simultaneously.

The deconvolution algorithms are highly non-linear, and the two methods are therefore not equivalent. The first one is straightforward to implement, but the non-linear mosaicing algorithms give much better results. Indeed, the combination of the adjacent fields in a mosaic allows to estimate visibilities which were not observed (see previous paragraph), it allows to remove sidelobes in the whole mapped area, and it increases the sensitivity in the (large) overlapping regions: these effects make the deconvolution much more efficient.

Non-linear deconvolution methods based on the `MEM` algorithm have been proposed by [Cornwell 1988]) and [Sault et al 1996]. As CLEAN deconvolutions are usually applied on Plateau de Bure data, a CLEAN-based method adapted to the case of the mosaics has been developed. The initial idea was proposed by F. Viallefond (DEMIRM, Paris) and S. Guilloteau (IRAM), and the algorithm is now implemented in the MAPPING software.

15.4 A CLEAN-based algorithm for mosaic deconvolution

The dirty mosaic

The dirty maps of each field i are computed with the same phase center (i.e. the same coordinate system) and can therefore be written:

$$F_i = D_i * (B_i \times I) + N_i \tag{15.10}$$

Note that the dirty beams D_i are *a priori* different for each pointing center, because the *uv* coverages, even if similar, are slightly different. The dirty mosaic J can then be constructed according to equation 15.9:

$$J = \frac{\sum_i \frac{B_i}{\sigma_i^2} F_i}{\sum_i \frac{B_i^2}{\sigma_i^2}} = \frac{\sum_i B_i \, \sigma_i^{-2} \left[D_i * (B_i \times I) + N_i \right]}{\sum_i B_i^2 \, \sigma_i^{-2}} \tag{15.11}$$

This relation is homogeneous to the sky brightness distribution I: the mosaic is corrected for the primary beams attenuation. In practice, a slightly modified mosaic is computed, in order to avoid noise propagation (it makes no sense to add to the center of a field noise coming from the external, attenuated regions of an adjacent field). For that purpose, the primary beams used to construct the mosaic are truncated to some value, typically 10 to 30% of the maximum. The mosaic is thus defined by:

$$J = \frac{\sum_i \frac{B_i^t}{\sigma_i^2} F_i}{\sum_i \frac{B_i^{t\,2}}{\sigma_i^2}} = \frac{\sum_i B_i^t \, \sigma_i^{-2} \left[D_i * (B_i \times I) + N_i \right]}{\sum_i B_i^{t\,2} \, \sigma_i^{-2}} \tag{15.12}$$

where B_i^t denotes the truncated primary beam of the field i. This relation is the measurement equation of a mosaic, connecting the observed quantity J to the sky brightness distribution I (equation 15.1 was the measurement equation of a single-field observation).

Noise distribution

Due to the correction for the primary beams attenuation, the noise distribution in a mosaic is not uniform. From Eq.15.12, it can be written:

$$N = \frac{\sum_i B_i^t \, \sigma_i^{-2} \, N_i}{\sum_i B_i^{t\,2} \, \sigma_i^{-2}} \tag{15.13}$$

Accordingly, the rms depends on the position and is given by:

$$\sigma_J = \frac{\sqrt{\sum_i B_i^{t\,2} \, \sigma_i^{-2}}}{\sum_i B_i^{t\,2} \, \sigma_i^{-2}} = \frac{1}{\sqrt{\sum_i B_i^{t\,2} \, \sigma_i^{-2}}} \tag{15.14}$$

The noise thus strongly increases at the edges of the mosaic (see Fig. 15.1). The non-uniformity of the noise level with the position makes it impossible to use classical CLEAN methods to deconvolve the mosaic: the risk to identify a noise peak as a CLEAN component would be too important. It is thus necessary to identify the CLEAN components on another distribution. For that purpose, the "signal-to-noise" distribution is computed:

$$H = \frac{J}{\sigma_J} = \frac{\sum_i B_i^t \, \sigma_i^{-2} \, F_i}{\sqrt{\sum_i B_i^{t\,2} \, \sigma_i^{-2}}}$$

$$\text{i.e.:} \quad H = \frac{\sum_i B_i^t \, \sigma_i^{-2} \left[D_i * (B_i \times I) + N_i \right]}{\sqrt{\sum_i B_i^{t\,2} \, \sigma_i^{-2}}} \tag{15.15}$$

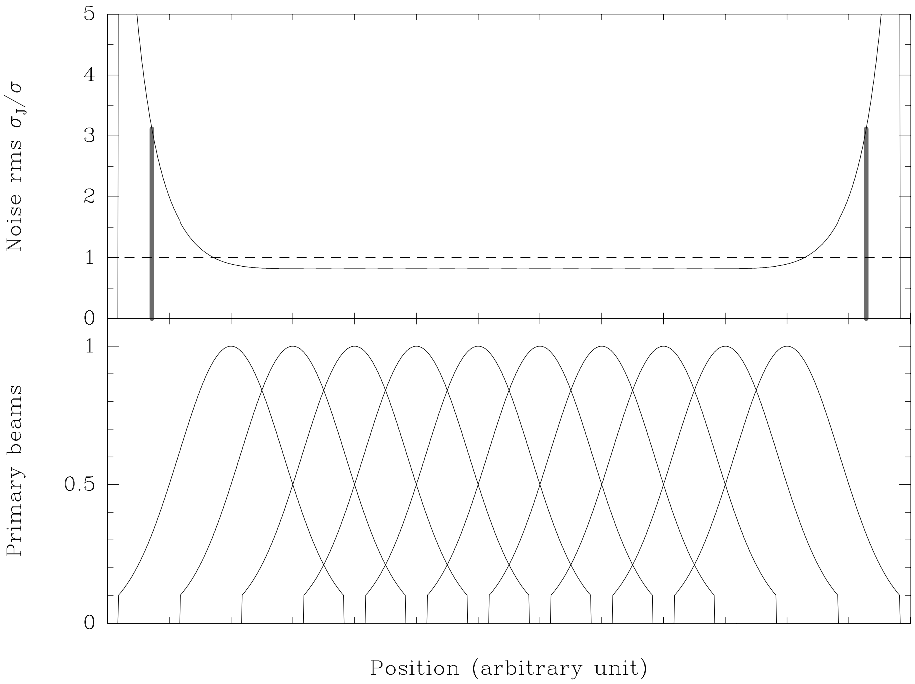


Figure 15.1: One-dimensional mosaic of 10 half-power overlapping fields, with identical noise level σ. (Lower panel:) Normalized primary beams, truncated to $\mathrm{B}_{\min} = 0.1$. (Upper panel:) Resulting noise distribution (Eq.15.14). The noise rms in the mosaic is roughly constant, about 20% lower than the noise of each individual field, but strongly increases at the edges. The two thick vertical lines indicate the truncation of the mosaic done by the algorithm at $\sigma_J = \sigma/\sqrt{\mathrm{B}_{\min}}$.

Deconvolution algorithm

The main idea of the algorithm is to iteratively find the positions of the CLEAN components on H, and then to correct the mosaic J. The initial distributions J_0 and H_0 are computed from the observations and the truncated primary beams, according to equations 15.12 and 15.15. The following operations have then to be performed at each iteration k:

1. Find the position (x_k, y_k) of the maximum of H.
2. Find the value j_k of J at the position (x_k, y_k).
3. Remove from J the contribution of a point-like source of intensity γj_k, located at (x_k, y_k) (γ is the loop gain, as in the normal CLEAN algorithm):

$$J_k = J_{k-1} - \frac{\sum_i B_i^t\,\sigma_i^{-2}\left[D_i * \left[\gamma\, j_k\, B_i(x_k, y_k)\,\delta(x_k, y_k)\right]\right]}{\sum_i {B_i^t}^2\,\sigma_i^{-2}} \tag{15.16}$$

   $\delta(x_k, y_k)$ denotes a Dirac peak located in (x_k, y_k).

4. Do the same for H: remove the contribution of a point-like source of intensity γj_k, located at (x_k, y_k):

$$H_k = H_{k-1} - \frac{\sum_i B_i^t\,\sigma_i^{-2}\left[D_i * \left[\gamma\, j_k\, B_i(x_k, y_k)\,\delta(x_k, y_k)\right]\right]}{\sqrt{\sum_i {B_i^t}^2\,\sigma_i^{-2}}} \tag{15.17}$$

   Note that in the two last relations, the CLEAN component is multiplied by the true, not truncated primary beam (taken at the (x_k, y_k) position).

After $k_{\max}$ iterations, the mosaic J can therefore be written:

$$J = \frac{\sum_i B_i^t\,\sigma_i^{-2}\left[D_i * \left(B_i \times \left[\sum_{k=1}^{k_{\max}} \gamma\, j_k\, \delta(x_k, y_k)\right]\right)\right]}{\sum_i {B_i^t}^2 \sigma_i^{-2}} + J_{k_{\max}} \tag{15.18}$$

Enough iterations have to be performed to ensure that the residual $H_{k_{\max}}$ is smaller than some user-specified threshold (typically 1 to 3). The comparison between Eqs.15.12 and 15.18 shows that, within the noise, the sum of the CLEAN components can be identified with the sky brightness distribution I. As with the normal CLEAN algorithm, the final clean image is then reconstructed as:

$$M = C * \left[\sum_{k=1}^{k_{\max}} \gamma\, j_k\, \delta(x_k, y_k)\right] + J_{k_{\max}} \tag{15.19}$$

where C is the chosen clean beam. The modified CLEAN algorithms proposed e.g. by [Clark 1980] or [Steer et al 1984] can be similarly adapted to handle mosaics, the main idea being to identify CLEAN components on H and to correct J. Note however that the multi-resolution CLEAN [Wakker & Schwartz 1988] cannot be directly adapted, as it relies on a linear measurement equation, which is not the case for a mosaic.

The MAPPING software

MAPPING is an superset of the GRAPHIC software, which has been developed to allow more sophisticated deconvolutions to be performed. For instance, it allows to choose a support for the deconvolution (clean window) or to monitor the results of the deconvolution after each iteration. Several enhancements of CLEAN (e.g. multi-resolution CLEAN) as well as the `WIPE` algorithm (see lecture by E. Anterrieu) are also available. The deconvolution of a mosaic has to be done with MAPPING. The implemented algorithm assumes that the noise levels in each field are similar (i.e. $\forall i\ \sigma_i = \sigma$), which is a reasonable hypothesis for Plateau de Bure observations. In that case, the equations of the previous paragraph are slightly simplified: J is independent from σ, and H can be written as the ratio H'/σ, where H' is independent from σ and is actually used to localize the CLEAN components.

We refer to the *Mapping Cookbook* for a description of the MAPPING software. In short, to deconvolve mosaics:

- Create a *uv* table for each observed field. Then, run the `UV_MAP` task to compute a dirty map and a dirty beam for each field, with the *same* phase center (variable `UV_SHIFT = YES`).
- The task `MAKE_MOSAIC` is used to combine the fields to construct a dirty mosaic. Two parameters have to be supplied: the width and the truncation level $B_{\min}$ of the primary beams. Three images are produced: the dirty mosaic[3] (yourfile.lmv), all the dirty beams written in the same file (yourfile.beam) and a file describing the positions and sizes of the primary beams (yourfile.lobe). The dirty maps and beams of each individual field are no longer used after this step and can thus be removed if necessary.
- The data have to be loaded into the MAPPING buffers. This is done by the `READ DIRTY yourfile.lmv`, `READ BEAM yourfile.beam`, and `READ PRIMARY yourfile.lobe` commands. The latter automatically switches on the mosaic mode of MAPPING (the prompt is now `MOSAIC>`). From now, the deconvolution commands `HOGBOM`, `CLARK` and `SDI` can be used and will apply the algorithm described above. Use the command `MOSAIC` to switch on or off the mosaic mode if necessary.
- The clean beam of the final image can be specified by the user (variables `MAJOR` and `MINOR`). Otherwise, the clean beam computed from the *first* field will be used. To check if there are differences between the various dirty beams, just use the `FIT` i command, which will indicates the clean beam computed for the ith field.

[3] More precisely, this file contains the non normalized mosaic $\Sigma B_i^t \times F_i$. The proper normalization (see equation 15.12) is further done by the deconvolution procedures.

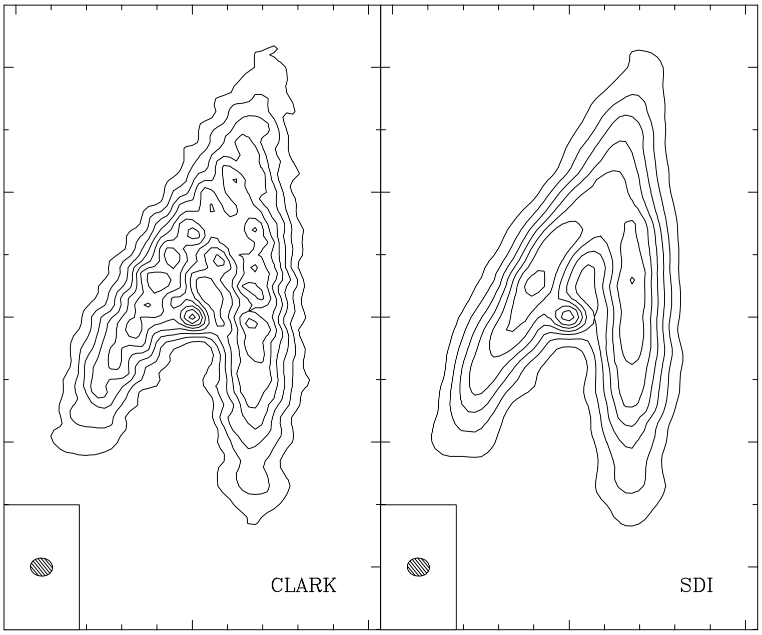


Figure 15.2: Mosaic deconvolved with the CLARK or SDI algorithms. Deconvolution parameters were identical (with a loop gain 1) and contours are the same in the two images. The formation of stripes does not occur when using the SDI algorithm.

- The deconvolution will use the same parameters as a usual CLEAN: support, loop gain, maximal number of iterations, maximal value of the final residual, etc.

- In addition, two other parameters, SEARCH_W and RESTORE_W, can be supplied. Due to the strong increase of the noise at its edges, the mosaic has to be truncated above some value of σ_J, and these two variables are used to define this truncation level (in terms of $(\sigma_J/\sigma)^{-2}$). More precisely, SEARCH_W indicates the limit above which CLEAN components have not to be searched, while RESTORE_W indicates the limit above which the clean image is not reconstructed. Default values of these two parameters (both equal to $B_{\min}$) are strongly recommended. The corresponding truncation is shown in Fig. 15.1.

Tests of the method

Several tests of the method described in this paragraph have been performed, either with observations (including the comparison of independent mosaics from the same source) or with simulations. They show that very satisfactory results can be achieved with typical Plateau de Bure observations. Interestingly, MEM deconvolution of the same data set (using the task VTESS in AIPS) seems to give worse results: this is most probably related to the limited uv-coverage obtained with the Plateau de Bure interferometer, as compared to typical VLA observations (MEM is known to be vulnerable when there is a relatively small number of visibilities).

15.5 Artifacts and instrumental effects

The behaviour of the mosaicing algorithm towards deconvolution artifacts and/or instrumental effects can be studied by the means of simulations of the whole mosaicing process. The models presented below were computed with several synthetic sky brightness distributions. uv coverage of real observations were used (4–antennas CD configuration of a source of declination $\delta = 68°$). No noise has been added to the simulations shown in the figures, so that pure instrumental effects can directly be seen.

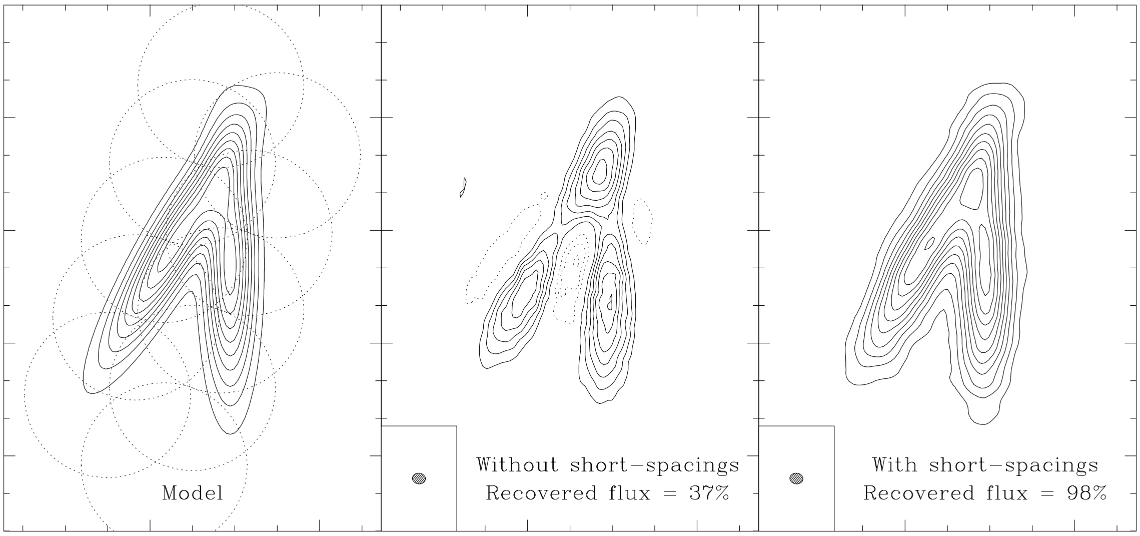


Figure 15.3: (Left:) Initial model of a very extended sky brightness distribution. Dotted circles indicate the primary beams of the simulated observation. (Middle:) Reconstructed mosaic, without the short-spacings information. (Right:) Reconstructed mosaic, with the short-spacings information. The contours are the same in the two simulated observations.

Stripes

A well-known instability of the CLEAN algorithm is the formation of stripes during the deconvolution of extended structures. After the dirty beam has been subtracted from the peak of a broad feature, the negative sidelobes of the beam are showing up as positive peaks. The next iterations of the algorithm will then identify these artificial peaks as CLEAN components. A regular separation between the CLEAN components is thereby introduced and the resulting map shows ripples or stripes. [Steer et al 1984] presented an enhancement of CLEAN (command `SDI` in MAPPING) which prevents such coherent errors: the CLEAN components are identified and removed in groups. As mosaics are precisely observed to map extended sources, the formation of stripes can *a priori* be expected. Indeed, the algorithm described in the previous paragraph presents this instability. Fig. 15.2 shows an example of the formation of such ripples. To make them appear so clearly, an unrealistic loop gain ($\gamma = 1$) was used. But the algorithm of [Steer et al 1984], adapted to the mosaics, does not result in these stripes, even with the same loop gain. It seems thus to be a very efficient solution to get rid of this problem, if it should occur. Note however that more realistic simulations, including noise and deconvolved with normal loop gain, do not show stripes formation. This kind of artifacts seems thus not to play a significant role in the image quality, for the noise and contrast range of typical Plateau de Bure observations. In practice, they are never observed.

Short spacings

The missing short spacings have potentially strong effects on the reconstructed brightness distributions in a mosaic. In each field, the most extended structures are filtered out, which thus introduces a lack of information on an *intermediate* scale as compared to the size of the mosaic. As a consequence, a very extended emission can be split into several pieces, each one having roughly the size of the primary beam. This effect can be very well seen on the simulation presented in Fig. 15.3. Should this problem occur, the only way to get rid of it is to add the short spacings information (deduced typically from single-dish observations) to the data set. Note however that the effects of the missing short-spacings on the reconstructed mosaic strongly depend on the actual uv coverage of the observations, as well as on the size and morphology of the source: the artifacts can be small or negligible if the observed emission is confined into reasonably small regions. From this point of view, the example shown in Fig. 15.3 represents the worst case.

In any case, CLEAN is known to be not optimal to deconvolve smooth, extended structures. In order

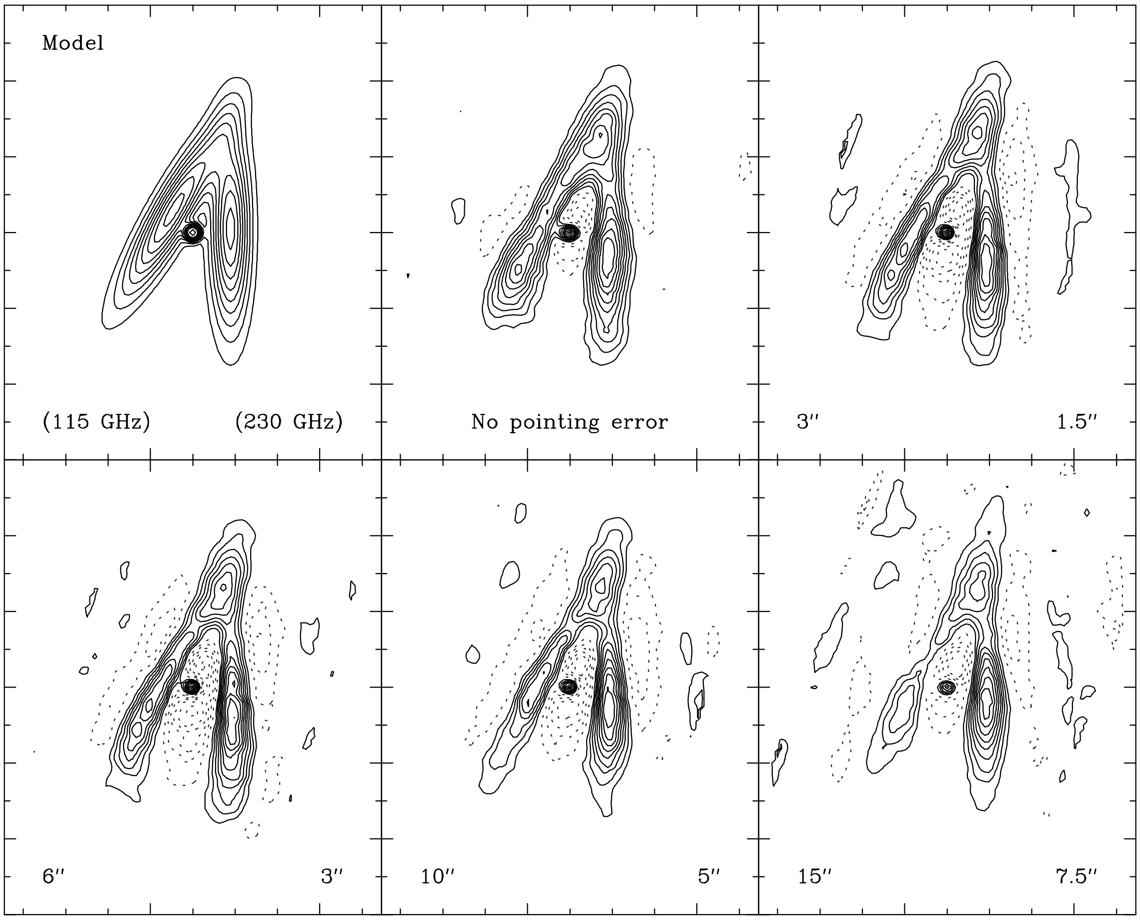


Figure 15.4: Simulations of a 10–fields mosaic observed with the Plateau de Bure interferometer. Each field is affected by a pointing error (see text). The corresponding rms are indicated in the lower left (observations performed at 115 GHz) and lower right (230 GHz) corners of each panel.

to partially alleviate this problem and the effects of the missing short-spacings, [Wakker & Schwartz 1988] proposed an enhanced algorithm, the so-called multi-resolution CLEAN: deconvolutions are performed at low- and high-resolution, and the results are combined to reconstruct an image which then accounts for the extended structures much better than in the case of a classical CLEAN deconvolution. As already quoted before, this algorithm cannot be applied to a mosaic, because it relies on a linear measurement equation. A multi-resolution CLEAN adapted to mosaics has however been developed ([Gueth 1997]) and is currently implemented in MAPPING. This method will not be described here.

Pointing errors

Pointing errors during the observations can of course strongly affect the images obtained by mosaicing. The rms of the pointing errors of the antennas of the Plateau de Bure interferometer is about $3''$. By comparison, the primary beam size at 230 GHz is $\sim 22''$ (Table 15.1). The pointing errors are difficult to model precisely: they are different for each antenna, random errors as well as slow drifts occur, the amplitude calibration partially corrects them, etc. A complete simulation should therefore introduce pointing errors during the calculation of each visibility. For typical Plateau de Bure observations, such a detailed modeling is probably not necessary, as the final image quality is dominated by deconvolution artifacts. To get a first guess of the influence of pointing errors, less realistic simulations were thus performed, in which each field is shifted as a whole by a (random) quantity. Such a systematic effect most probably maximizes the distortions introduced in the images. (Note that for a single field, the source would simply be observed at a shifted position in such a simulation. For a mosaic, the artifacts are different, as each individual field has a different, random pointing error. See [Cornwell 1987] for a simplified analysis in terms of visibilities.) Figure 15.4 presents typical reconstructed mosaics for different rms of the pointing errors of the Plateau de Bure antennas. Obviously, the larger the pointing error, the worse the image quality. With a pointing error rms of $3''$, reasonably correct mosaics can be reconstructed even at 230 GHz. Clearly, care to the pointing accuracy has however to be exercised when mosaicing at the highest frequencies.

15.6 Concluding remarks

Mosaic observations are now routinely performed with the Plateau de Bure interferometer, at both λ 3 mm and λ 1.3 mm. Data processing requires a few more operations than normal observations, but does not present any specific difficulties. Reconstruction and deconvolution algorithms are available in the MAPPING software. The number of mosaics actually observed with the Plateau de Bure interferometer regularly increased during the last years. For the last observing period, it amounts about 40% of the mapping projects (the fraction of observing time used for mosaics is much lower, as many time-consuming detection projects are also performed). The number of fields are usually ≤ 6, but can be more important in some cases: the largest mosaic observed up to now (September 1998) has 13 fields.

Chapter 16

Imaging in Practice

Stéphane Guilloteau

IRAM, 300 rue de la Piscine, F-38406 Saint Martin d'Hères

16.1 Visualisation

Contrary to the lower frequencies where continuum emission processes are dominant, mm interferometry most frequently deals with spectral lines, and hence involves handling and display of *data cubes.* Only a few astronomy packages have been designed for this: GIPSY, GILDAS, MIRIAD. Although the presentation which follows is general enough about the principles (e.g. for the noise analysis or flux density measurements), I will only present the tools which are currently available in GILDAS. Within GILDAS, two display tools are available:
- The GRAPHIC program
- The MVIEW task

In GRAPHIC, easy display is available using the following commands
- `GO MAP`, for simple channel contour maps
- `GO BIT`, same with overlaid color bit map
- `GO NICE`, with clean beam in addition
- `GO POS`, for Position-Velocity plots
- `GO SPECTRE`, for maps of spectra.

Easy access to the parameters of these procedures is available through the Windowing interface. GRAPHIC also provides access to all image processing tasks such as `UV_MAP`, `UV_STAT`, `CLEAN`, etc..., and flexible controls for publication quality plots.

Task MVIEW provides a different approach. It is a Window based application which provides simple, intuitive, and fast interactive 3-D data cube display. It provides spectrum display at cursor position, slices, moments, movie features, color manipulation, etc ... It also has a direct interface to some important tasks (e.g. moments evaluation, subset extraction) for which interactively selecting parameters using mouse motion is convenient. However, contrary to GRAPHIC, it cannot be customized to produce publication quality images.

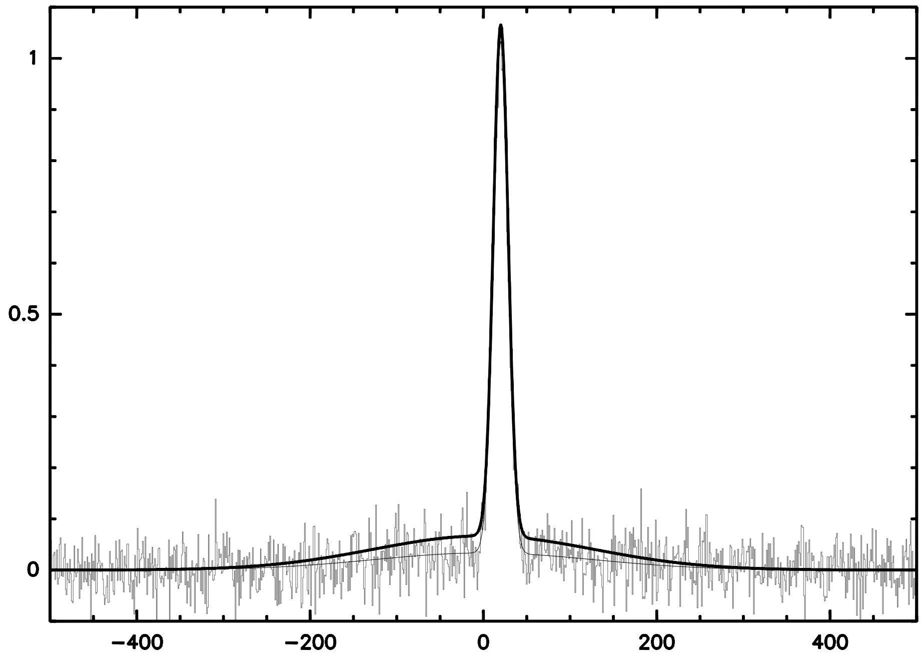


Figure 16.1: Illustration of the difficulty to deconvolve a weak, extended structure. In this 1-D example, half of the flux is in the extended structure, and cannot be recovered properly by deconvolution because of low signal to noise.

16.2 Photometry

16.2.1 From Flux density to Brightness temperature

The unit of the dirty map is ill defined. A single point source of 1 Jy appears with peak intensity of 1. But if more than 1 point source is in the field of view, the combination of positive or negative sidelobes from the other source modify this result. It is thus necessary to deconvolve. After deconvolution, the beam area is well defined: the CLEAN map unit is Jy per beam area.

The conversion to brightness temperature can then be done using the standard equation

$$S_\nu = \frac{2k\Omega_s}{\lambda^2} T_B \tag{16.1}$$

$$= \frac{2k\pi\theta_s^2}{4\log 2\lambda^2} T_B \tag{16.2}$$

for Gaussian beams. `GO MAP` and its variants automatically display the Jy/K conversion factor mentioned above. The integrated flux density in a user defined area can then be computed on the clean map using command `GO FLUX`.

16.2.2 Accuracy of Flux density estimates

The accuracy of flux measurements is limited by several factors.

Deconvolution Errors and Missing Flux

Deconvolution limits are among the most important. Deconvolution is required (see Chapter 13), but it is impossible to deconvolve weak structures near the noise level. Nevertheless, these structures, when sufficiently extended, can contribute to a significant flux. The relevance of the missing flux to the astronomical interpretation is to be decided by the astronomer. In most cases, however, the missing flux does not correspond to any significant brightness, and its absence may not modify the astronomical interpretation. A schematic illustration is given in Fig.16.1: half of the flux is buried in the noise, but the brightness is measured properly (within the statistical error). Note also that WIPE offers an upper limit to the noise amplification factor due to the non linear deconvolution process. This upper limit is often discouragingly large for the rather poor dirty beams provided by current mm arrays (4 or 5 is not unusual).

Comparing the recovered flux to a single dish measurement can help you in estimating the corresponding brightness level and evaluate whether this is important in your astronomical case. To convert from flux to brightness, the characteristic size of the missed flux should be used. This is in between several synthesized beam widths and about half of the primary beam. Although in general the problem is not as severe as one would think (because, unfortunately, one is often signal to noise limited), this is an important information, specially in the case of mosaics.

Seeing

A second important effect which affects flux density measurements is the seeing. Seeing result in an underestimation of Point source flux. On the other hand, the total flux is spread over the seeing disk, and is in principle conserved. Insufficient seeing conditions limit the deconvolution process, since the effective synthesized beam (which should include atmospheric phase errors) is significantly different from the theoretical synthesized beam (computed from *uv* coverage and weights). A good check is to make an image of the two calibrators, and measure the corresponding point source flux and apparent size.

Noise estimate

Finally, the noise should be estimated. `GO RMS` gives the sigma of the image flux density distribution. This is an improper estimate, since it includes any possible signal, all spectral channels, and map edges where the noise level increases due to aliasing and gridding. A better estimate, derived from effective weights, is in principle given by task `UV_STAT`. However, this estimate does not take into account possible deconvolution problems or dynamic range limitations due to atmospheric phase noise. The optimal procedure is to use command `GO FLUX` on an empty area of the image to find out the point source rms noise. The precision of the estimate is limited by statistical uncertainties linked to the number of beams in the area. Then, another `GO FLUX` command on the emission area will give the total flux and number of independent beams, n; the rms on the total flux is $\sqrt{n}$ times the point source rms determined by the `GO FLUX` command applied to an empty region. Another good method to determine the noise level (yet to be implemented as a `GO NOISE` procedure ...) would be to build the histogram of the pixel values and fit a Gaussian to it; if source structure only covers a small fraction of the image, this method provides a good estimate.

Primary beam

One should emphasize that primary beam correction is essential in any correct flux density estimate. All image plane analysis should be carried out on a primary beam corrected image. This introduces a slight complication, since the noise level is then not uniform. The `GO FLUX` commands discussed before should be applied in regions of similar extent and location vis-a-vis the primary beam(s).

uv plane analysis is also extremely useful both in measuring integrated flux densities and rms noise level, at least for simple, relatively compact, source models. Task `UV_FIT` provides statistical errors for all parameters of the fit. Primary beam correction should be applied a posteriori, based on the location of the region of interest.

Dynamic Range

As mentioned above, the dynamic range may be a limitation. The dynamic range D is defined as the ratio of the peak intensity to the lowest “believable” contour. D is obviously lower than the signal to noise ratio. It can be estimated as the absolute value of the peak to maximum negative contour ratio. As usual, map edges should not be included in this evaluation. Dynamic range is related to seeing and calibration errors. It is typically 10 to 40 at Plateau de Bure (if signal to noise ratio allows). Errors should include dynamic range effects.

Flux density scale

Finally, remember that the flux scale is determined by bootstrapping flux of (variable) quasars from that of reference sources. Any errors accumulated in this process must be transferred to the source flux estimate.

In summary, flux density estimates should quote errors which include

- Effective thermal noise
- Dynamic range problems
- Relative (calibrator) flux uncertainty
- Absolute flux scale uncertainty **OR** reference flux scale.
- Primary beam correction

16.3 Short Spacings

Extended structure are missed, attenuated or distorted in interferometric maps by lack of short *uv* spacing information. While this effect may be negligible for some astronomical problems, it could also be essential in a proper analysis. Deconvolution recovers some of them, but under-estimate the total flux because the integral of the dirty beam is zero (the integral of the dirty beam is the weight of the (0,0) *uv* cell in the *uv* data set).

Constructing a beam with a non zero integral can help deconvolution. This can be done by incorporating the **Zero spacing** flux or spectrum.

Short spacings provides even more information, because they give information on the spatial distribution of this flux on scales between half the primary beam and the primary beam itself. Short spacings can be provided by a smaller interferometer (e.g. BIMA) or a large single dish (e.g. 30-m). In theory, short spacings can also be provided by the interferometer antennas used in single-dish mode. However, because most interferometer have not been designed with total power stability as a goal, this has not been practiced so far.

Incorporating short spacings into interferometer data is a two step process. Task `UV_SINGLE` extracts short spacing information from single dish data (spectra) and creates a *uv* table. Task `UV_MERGE` merges the single-dish and interferometer tables. Coordinates system should be consistent and checked before (coordinates are always J2000.0 at Plateau de Bure, often B1950.0 at the other observatories...).

16.3.1 UV_SINGLE

Incorporating short spacings from the 30-m into Plateau de Bure data is a 3 step process for the user.

- Creation of a table of spectra
  First, one should resample (in frequency) all spectra to same frequency grid than interferometer data, using command `RESAMPLE` in CLASS). Then a table of spectra is produced using command `GRID` (with no options) in CLASS.

- Image creation
  The next step is the creation of "well behaved" map from the table of spectra. It starts with resampling (in space) on a regular grid by a convolution kernel (interpolation techniques such as the GREG command `RANDOM_MAP` are inappropriate). Weights are also resampled. Then, we follow by extrapolation to zero outside the convex hull of the mapped region, using a kernel twice broader than the single-dish beam, to avoid introducing spurious structure.

  Because of noise, the map still contains spurious high spatial frequencies. These are removed during the *uv* table creation. The algorithm steps are

  - Fourier transform of map and weight images

  - Division by Fourier Transform of the single-dish beam
  - Gridding correction (division by Fourier Transform of the gridding function)
  - Truncation to some maximum *uv* distance (< dish diameter)
  - Inverse Fourier Transform back to image plane
  - Multiplication by primary beam of the interferometer
  - Fourier Transform to *uv* plane
  - Normalization of the weights so that the sum of weights is the weight of the total flux (derived from integration time, bandwidth and system temperature).
  - Optional application of an amplitude scaling factor.
  - Optional application of a weight scaling factor.

  This produces a *uv* table with optimal weights in terms of signal to noise ratio for the total flux, and with effective tapering following the single-dish illumination pattern, except for the truncation at some *uv* distance.

- Merging with interferometer data
  The final step is to merge the resulting *uv* table with the interferometer *uv* table. Re-weighting and re-scaling is again possible at this stage. Note that the choice of weighting function is arbitrary. It may result in poorly behaved synthesized beams when combined with the interferometer *uv* data. Weights can be lowered by any arbitrary factor (increasing the weights is only allowed if signal to noise is not an issue). A good choice is to adjust the weights so that there is almost no negative sidelobe.

16.4 Dirty Tricks

Besides flux density estimate, which, as discussed before, is a non trivial task, analyzing spectral line images may force the astronomer to face some really tricky problems. The two most obvious are moment evaluation and continuum subtraction.

16.4.1 MOMENTS

The lowest order moment of a spectral line data cubes offer very convenient ways of interpreting images. The zero$^{\text{th}}$ order moment is the integrated intensity, the first order moment the velocity, the second order moment the line width. While these moments are linear combination of the channel maps, the deconvolution process is non linear. Accordingly, the two operations do not commute.

Hence, it is impossible to recommend deconvolving before computing the mean intensity, or summing up the individual cleaned channel maps. In the latter, limited signal to noise can prevent proper deconvolution. In the former, velocity gradients can spread emission over an extended area which is difficult to handle in the deconvolution. Choice can be a matter of trial (and errors).

To avoid introducing noise, a *window* in velocity is important. While noise on the integrated intensity only increases as the square root of the window width, the effect on the higher order moments is much more dramatic, and results in non-gaussian noise distribution on these variables. A *threshold* in intensity is useful to prevent spurious noisy features. The window should in principle be pixel dependent to allow for velocity gradients. Smoothing both in the spatial and spectral domains may help in obtaining better results in moment extraction. A line fitting procedure (e.g. a Gaussian line fit at each pixel) may sometimes be the best solution (under construction, check later...).

Moments can be computed using task `MOMENTS` and displayed using the `GO VELOCITY` command in GRAPHIC.

16.4.2 Continuum Subtraction

Continuum subtraction is a related problem. It is in principle needed to compute properly moment maps. However, it may be completely impossible, for example in the case of an optically thick line partially covering a continuum source. Continuum subtraction can be done in the image plane or in the *uv* plane. *uv* plane subtraction avoid the non linearity in the deconvolution, and thereby any amplification of errors induced in this process. Task `UV_SUBTRACT` performs this operation. Although signal to noise on the continuum is often much better than on the spectral line, it may be advantageous to subtract a source model rather than the measured visibilities; this is only true when thermal noise is more important than phase noise. Task `UV_MODEL` compute visibilities from an input image.

Chapter 17

Basic Principles of Radio Astrometry

Alain Baudry

Observatoire de Bordeaux, BP 89, F-33270 Floirac

17.1 Introduction and Basic Formalism

Modern astrometry aims at improving our knowledge of celestial body positions, motions and distances to a high accuracy. The quest for accuracy began in the early days of astronomy and is still continuing in the optical domain with most sophisticated instruments (automated meridian circles, the Hipparcos satellite or future astrometric space missions) as well as in the radio domain (connected-element interferometers and VLBI). New instrumental concepts or calibration procedures and increased sensitivity are essential to measure highly accurate positions of stars and radio sources. Positions accurate to about one thousandth to one tenth of an arcsecond have now been obtained for hundreds of radio sources and for about 100 000 to one million stars in the Hipparcos and Tycho catalogues respectively.

In this lecture we are concerned with some basic principles of position measurements made with synthesis radio telescopes and with the IRAM interferometer in particular. More details on interferometer techniques can be found in the fundamental book of [Thompson et al 1986]. The impact of VLBI in astrometry and geodesy is not discussed here. (For VLBI techniques see [Sovers et al 1998])

We first recall that measuring a position is a minimum prerequisite to the understanding of the physics of many objects. One example may be given for illustration. To valuably discuss the excitation of compact or masing molecular line sources observed in the direction of late-type stars and HII regions sub-arcsecond position measurements are required. This is because the inner layers of circumstellar envelopes around late-type stars have sizes of order one arcsecond or less and because several compact HII regions have sizes of one to a few arcseconds only. Position information is crucial to discuss not only the respective importance of radiative and collisional pumping in these line sources but also the physical asssociation with the underlying central object.

The output of an interferometer per unit bandwidth at the observing wavelength λ is proportional to the quantity

$$R = \int A(\vec{k}) I(\vec{k}) \cos(2\pi \vec{B}.\vec{k}/\lambda) d\Omega \tag{17.1}$$

where $\vec{k}$ is the unit vector toward the observed source, A is the effective antenna aperture, I the source brightness, and $\vec{B}$ the baseline vector of the interferometer. For an extended source one refers the observations to the reference direction $\vec{k}_o$ and supposing that the radiation comes from a small portion of the sky we have $\vec{k} = \vec{k}_o + \vec{\sigma}$ where $\vec{\sigma}$ is the position vector describing the source coordinates. (Since both $\vec{k}_o$ and $\vec{k}$ are unit vectors we obtain $\vec{k}_o.\vec{\sigma} = 0$.) The interferometer output is given by

$$R = V\cos(2\pi\vec{B}.\vec{k}_o/\lambda + \Psi) \tag{17.2}$$

where

$$V\exp(i\Psi) = \int A(\vec{\sigma})I(\vec{\sigma})\exp(i2\pi\vec{b}.\vec{\sigma})d\Omega \tag{17.3}$$

is the complex source visibility and $\vec{b}(u,v)$ is the baseline vector projected on a plane normal to the tracked direction. The exact definition of the baseline coordinates u and v is given in Section 17.3.

The astrometry domain corresponds to those cases where the source visibility amplitude is equal to 1 (point-like sources) and the phase provides the source position information.

17.2 The Phase Equation

The most important measurement for radio astrometry is that of the actual fringe phase of a connected-element interferometer (or similarly the group delay in VLBI). Let θ be the angle between the reference direction and the meridian plane of a given interferometer baseline. The phase is then defined by

$$\phi_r = 2\pi B\sin(\theta)/\lambda \tag{17.4}$$

If the point-like source of interest is offset by $\Delta\theta$ from the reference direction the total phase is

$$\phi = 2\pi B\sin(\theta + \Delta\theta)/\lambda \simeq \phi_r + 2\pi B\cos(\theta)\Delta\theta/\lambda \tag{17.5}$$

It is thus clear that measuring an angle or an offset position on the celestial sphere becomes possible only when all phase calibration problems have been understood and solved.

Accounting for uncertainties in the baseline and source position vectors the actual phase is

$$\phi = 2\pi(\vec{B} + \delta\vec{B}).(\vec{k}_o + \delta\vec{k})/\lambda \tag{17.6}$$

where $\vec{B}$ is a first approximation of the baseline, $\vec{k}_o$ the tracking direction; $\vec{B} + \delta\vec{B}$ and $\vec{k}_o + \delta\vec{k}$ are the true baseline and source position vectors, respectively. The reference phase is given by

$$\phi_r = 2\pi\vec{B}.\vec{k}_o/\lambda \tag{17.7}$$

and, neglecting the term involving $\delta\vec{B}.\delta\vec{k}$, we obtain

$$\phi - \phi_r = 2\pi(\vec{B}.\delta\vec{k} + \delta\vec{B}.\vec{k}_0)/\lambda \tag{17.8}$$

We consider all vector projections in the right-handed equatorial system defined by the unit vectors a_1 ($H =$ 6 h, $\delta = 0$), a_2 ($H = 0$ h, $\delta = 0$), a_3 ($\delta = 90°$). H and δ are the hour angle and declination, respectively. In this coordinate system the baseline vector $\vec{B}$ has components (B_1, B_2, B_3) and the components of the reference position $\vec{k}_o$ are given by $(\cos(\delta)\sin(H), \cos(\delta)\cos(H), \sin(\delta))$

The two limiting cases $\delta\vec{k} = 0$, and $\delta\vec{B} = 0$ correspond to those where we either calibrate the baseline or determine the exact source position.

In the first case the source coordinates are perfectly known and by comparing the observed phase ϕ with the reference phase ϕ_r one determines $\delta\vec{B}$ and hence the true baseline $\vec{B} + \delta\vec{B}$. The reference sources observed for baseline calibration are bright quasars or galactic nuclei whose absolute coordinates are accurately known. The most highly accurate source coordinates are those of the radio sources used to realize by VLBI the International Celestial Reference Frame (ICRF); distribution of coordinate errors are below one milliarcsecond. However, the ICRF catalogue is insufficient for phase and baseline calibrations of millimeter-wave arrays because most sources are not bright enough in the millimeter-wave domain. The IRAM calibration source list is thus a combination of several catalogues of compact radio sources.

17.3 Determination of Source Coordinates and Errors

Once the baseline is fully calibrated ($\delta\vec{B} = 0$) the exact source coordinates are known from the $\delta\vec{k}$ vector components. These components are formally deduced from the differential of $\vec{k}_o$

$$\begin{aligned}\delta\vec{k} \quad = \quad & (-\sin(\delta)\sin(H)\Delta\delta - \cos(\delta)\cos(H)\Delta\alpha, \\ & -\sin(\delta)\cos(H)\Delta\delta + \cos(\delta)\sin(H)\Delta\alpha, \\ & \cos(\delta)\Delta\delta)\end{aligned} \tag{17.9}$$

where $\Delta\alpha$ and $\Delta\delta$ are the right ascension and declination offsets in the equatorial system ($\Delta\alpha = -\Delta H$). The phase difference is then a sinusoid in H

$$\frac{(\phi - \phi_r)\lambda}{2\pi} = \vec{B}\delta\vec{k} = A\sin(H) + B\cos(H) + C \tag{17.10}$$

where

$$A \quad = \quad -B_1\sin(\delta)\Delta\delta + B_2\cos(\delta)\Delta\alpha \tag{17.11}$$
$$B \quad = \quad -B_2\sin(\delta)\Delta\delta - B_1\cos(\delta)\Delta\alpha \tag{17.12}$$
$$C \quad = \quad B_3\cos(\delta)\Delta\delta + \phi_{\rm ins} \tag{17.13}$$

and C contains the instrumental phase $\phi_{\rm ins}$.

Measurement of the phase at time intervals spanning a broad hour angle interval allows us to determine the three unknowns A, B, and C, and hence $\Delta\alpha$ and $\Delta\delta$ and the exact source position. Note that for sources close to the equator, A and B alone cannot accurately give $\Delta\delta$. In the latter case, C must be determined in order to obtain $\Delta\delta$; this requires to accurately know the instrumental phase and that the baseline is not strictly oriented along the E-W direction (in which case there is no polar baseline component).

A synthesis array with several, well calibrated, baseline orientations is thus a powerful instrument to determine $\delta\vec{k}$. In practice, a least-squares analysis is used to derive the unknowns $\Delta\alpha$ and $\Delta\delta$ from the measurements of many observed phases ϕ_i (at hour angle H_i) relative to the expected phase ϕ_r. This is obtained by minimizing the quantity
$\Sigma(\Delta\phi'_i - (A\sin(H_i) + B\cos(H_i) + C))^2$ with respect to A , B, and C where $\Delta\phi'_i = (\phi_i - \phi_r)\lambda/2\pi$. A complete analysis should give the variance of the derived quantities $\Delta\alpha$ and $\Delta\delta$ as well as the correlation coefficient.

Of course we could solve for the exact source coordinates and baseline components simultaneously. However, measuring the baseline components requires to observe several quasars widely separated on the sky. At mm wavelengths where atmospheric phase noise is dominant this is best done in a rather short observing session whereas the source position measurements of often weak sources are better determined with long hour angle coverage. This is why baseline calibration is usually made in separate sessions with mm-wave connected-element arrays.

The equation giving the source coordinates can be reformulated in a more compact manner by using the components u and v of the baseline projected in a plane normal to the reference direction. With v directed toward the north and u toward the east, the phase difference is given by

$$(\phi - \phi_r) = 2\pi(u\cos(\delta)\Delta\alpha + v\Delta\delta) \tag{17.14}$$

Comparing this formulation to the sinusoidal form of the phase difference we obtain

$$u \quad = \quad (-B_1\cos(H) + B_2\sin(H))/\lambda \tag{17.15}$$
$$v \quad = \quad (B_3\cos(\delta) - \sin(\delta)(B_1\sin(H) + B_2\cos(H)))/\lambda \tag{17.16}$$

Transforming the $B_{1,2,3}$ into a system where the baseline is defined by its length $B = (B_1^2+B_2^2+B_3^2)^{0.5}$ and the declination d and hour angle h of the baseline vector (defined as intersecting the northern hemisphere) we obtain

$$B_1 = B\cos(d)\sin(h), B_2 = B\cos(d)\cos(h), B_3 = B\sin(d) \tag{17.17}$$

and

$$\begin{aligned} u &= (\cos(d)\sin(H-h))B/\lambda \\ v &= (\cos(\delta)\sin(d) - sin(\delta)\cos(d)\cos(H-h))B/\lambda \end{aligned} \quad (17.18)$$

which shows that the locus of the projected baseline vector is an ellipse.

In order to derive the unknowns $\Delta\alpha$ and $\Delta\delta$ the least-squares analysis is now performed using the components u_i, v_i derived at hour angle H_i. In the interesting case where the phase noise of each phase sample is constant (this occurs when the thermal noise dominates and when the atmospheric phase noise is "frozen") one can show that the error in the coordinates takes a simple form. For a single baseline and for relatively high declination sources the position error is of order $\sigma_\varphi/(2\pi\sqrt{n_p}(B/\lambda))$ where σ_φ is the phase noise and n_p the number of individual phase measurements. This result implies (as expected a priori) that lower formal uncertainties are obtained with longer observing times and narrower synthesized beams. Of course the position measurements are improved with several independent interferometer baselines; the precision improves as the inverse of the square root of $n(n-1)/2$ for n antennas in the array.

We have shown that for a well calibrated interferometer the least-squares fit analysis of the phase in the (u, v) plane can give accurate source coordinates. However, the exact source position could also be obtained in the Fourier transform plane by searching for the coordinates of the maximum brightness temperature in the source map. The results given by this method should of course be identical to those obtained in the (u, v) plane although the sensitivity to the data noise can be different.

Finally, it is interesting to remind that the polar component of the baseline does not appear in the equation of the fringe frequency which is deduced from the time derivative of the phase. There is thus less information in the fringe frequency than in the phase.

17.4 Accurate Position Measurements with the IRAM Interferometer

Let us start with two general and simple remarks. First, the phase equation giving the angular offset θ in Sec.17.2 shows that higher position accuracy (namely smaller values of the angular offset) is achieved for smaller values of the fringe spacing λ/B. (This was demonstrated above in the case of the least squares analysis of the u, v data.) Thus, for astrometry it is desirable to use long baselines and/or to go to short wavelengths. However, the latter case implies that the phases are more difficult to calibrate especially at mm wavelengths where the atmospheric phase fluctuations increase with long baselines. Sensitivity is always important in radio astrometry. For a point-like or compact source the sensitivity of the array varies directly as $D^2(n(n-1))^{0.5}$ where D is the antenna diameter and n is the number of antennas. Thus the detection speed varies as $D^4 n(n-1)$ and big antennas are clearly advantageous. Comparison of the IRAM 5-element array with one of its competitors, OVRO with 6 x 10.4 m, gives a ratio of detection speed of 1 over 0.35 in favor of the Plateau de Bure array. (Note also that the sixth antenna in the Bure array will increase the detection speed by 50%.) In addition, the large dishes of the IRAM array are good to perform quick baseline and phase calibrations; this is another clear advantage of the IRAM interferometer in astrometric observations.

17.4.1 Absolute positions

To illustrate the potential of the IRAM array for astrometry we consider here observations of the SiO maser emission associated with evolved late-type stars. Strong maser line sources are excited in the $v = 1, J = 2-1$ transition of SiO at 86 GHz and easily observed with the sensitive IRAM array. Because of molecular energetic requirements (the vibrational state $v = 1$ lies some 2000 K above the ground-state) the SiO molecules must not be located too much above the stellar photosphere. In addition, we know that the inner layers of the shell expanding around the central star have sizes of order one arcsecond or less. Therefore, sub-arcsecond position accuracy is required to locate the SiO sources with respect to the underlying star whose apparent diameter is of order 20-50 milliarcseconds. For absolute position measurements one must primarily:

- select long baselines to synthesize small beamwidths,
- make a highly accurate baseline calibration observing several quasars selected for their small position errors,
- observe at regular intervals two or more quasars (phase calibrators) in the field of each program star in order to determine the instrumental phase and to correct for atmospheric phase fluctuations,
- observe the program star over a long hour angle interval, and use the best estimate of the stellar coordinates.

Our first accurate radio position measurements of SiO masers in stars and Orion were performed with the IRAM array in 1991/1992. We outline below some important features of these observations [Baudry et al 1994]. We used the longest E-W baseline available at that time, about 300 m, thus achieving beams of order 1.5 to 2 arcseconds. The RF and IF bandpass calibrations were made accurately using strong quasars only. To monitor the variable atmosphere above the array and to test the overall phase stability, we observed a minimum of 2 to 3 nearby phase calibrators. Prior to the source position analysis we determined accurate baseline components; for the longest baselines the r.m.s. uncertainties were in the range 0.1 to 0.3 mm. The positions were obtained from least-square fits to the imaginary part of the calibrated visibilities. (Note that the SiO sources being strong, working in the (u, v) or image planes is equivalent.)

The final position measurement accuracy must include all known sources of uncertainties. We begin with the formal errors related to the data noise. This is due to finite signal to noise ratio (depending of course on the source strength, the total observing time and the general quality of the data); poorly calibrated instrumental phases may also play a role. In our observations of 1991/1992 the formal errors were around 10 to 30 milliarcseconds. Secondly, phase errors arise in proportion with the baseline error $\delta\vec{B}$ and the offset between the unit vectors pointing toward the stellar source and the nearby phase calibrator. This phase error is $\delta(\phi - \phi_r) = (\delta\vec{B}.(\vec{k}_{quasar} - \vec{k}_o))2\pi/\lambda$. Typical values are $\delta B \simeq 0.2$ mm and $\delta k \sim 10° - 20°$ corresponding to phase errors of 3° to 7°, that is to say less than the typical baseline residual phases. A third type of error is introduced by the position uncertainties of the calibrators. This is not important here because the accuracy of the quasar coordinates used during the observations were at the level of one milliarcsecond.

The quadratic addition of all known or measured errors is estimated to be around $0.07''$ to $0.10''$. In fact, to be conservative in our estimate of the position accuracy we measured the positions of nearby quasars using another quasar in the stellar field as the phase calibrator. The position offsets were around $0.1''$ to $0.2''$ depending on the observed stellar fields; we adopted $0.1''$ to 0.2" as our final position accuracy of SiO sources. The SiO source coordinates are derived with respect to baseline vectors calibrated against distant quasars. They are thus determined in the quasi-inertial reference frame formed by these quasars.

Finally, it is interesting to remind a useful rule of thumb which one can use for astrometry-type projects with any connected-element array provided that the baselines are well calibrated and the instrumental phase is stable. The position accuracy we may expect from a radio interferometer is of the order of 1/10th of the synthesized beam (1/20th if we are optimistic). This applies to millimeter-wave arrays when the atmospheric fluctuations are well monitored and understood. With baseline lengths around 400 m the IRAM array cannot provide position uncertainties much better than about $0.05 - 0.1''$ at 86 GHz. Extensions to one kilometer would be necessary to obtain a significant progress; the absolute position measurements could then be at the level of 50 milliarcseconds which is the accuracy reached by the best optical meridian circles.

17.4.2 Relative Positions and Self-calibration Techniques

We have measured with the IRAM array the absolute position of the SiO emission sources associated with each spectral channel across the entire SiO emission profile. Any spatial structure related to the profile implies different position offsets in the direction of the star. Such a structure with total extent of about 50 milliarcseconds is observed in several late-type stars. This is confirmed by recent VLBI observations of SiO emission in a few stars. VLBI offers very high spatial resolution but poor absolute position measurements in line observations.

The best way to map the relative spatial structure of the SiO emission is to use the phase of one reference feature to map all other features. This spectral self-calibration technique is accurate because all frequency-independent terms are cancelled out. The terms related to the baseline or instrumental phase uncertainties as well as uncalibrated atmospheric effects are similar for all spectral channels and cancel out in channel to channel phase differences. By making the difference

$$(\phi(\nu) - \phi(\nu_{\rm ref}))(\lambda/2\pi) = (\vec{B}.\delta\vec{k}(\nu) - \vec{B}.\delta\vec{k}(\nu_{ref}) \tag{17.19}$$

where the SiO reference channel is at frequency ν_{ref} we obtain a phase difference equation whose solution gives the coordinate offsets $\Delta\alpha(\nu)$ and $\Delta\delta(\nu)$ relative to channel ν_{ref}. The main limitation in such self-calibration techniques comes from the thermal noise and the achieved signal to noise ratio SNR. The angular uncertainty $\Delta\theta$ can then be estimated with the simple equation $\Delta\theta = 0.5(\lambda/B)/\mathrm{SNR}$. Common practice with connected-element arrays shows that selection of a reference channel is not critical; it must be strong in general. Self-calibration proved to be successful with the IRAM array in several stars and in Orion where we have obtained very detailed relative maps of SiO emission. Detailed relative maps were also obtained for the rare isotope ^{29}SiO; this emission is nearly 2 orders of magnitude weaker than that of the main isotope.

The relative spot maps obtained with connected-element arrays do not give the detailed spatial extent of each individual channel. This would require a spatial resolution of about one milliarcsecond which can only be achieved with VLBI techniques. Note however that VLBI is sensitive to strong emission features while the IRAM array allows detection of very weak emission; thus the two techniques appear to be complementary.

With SiO spatial extents of about 50 milliarcseconds and absolute positions at the level of 0.1 arcsecond it is still difficult to locate the underlying star. We have thus attempted to obtain simultaneously the position of one strong SiO feature relative to the stellar photosphere and the relative positions of the SiO sources using the 1 and 3 mm receivers of the IRAM array. This new dual frequency self-calibration technique is still experimental but seems promising.

Bibliography

[Altenoff et al 1994] Altenhoff, W., Thum, C., Wendker, H.J., 1994, *A&A* **281**, 161-183
Radio emission from stars: a survey at 250 GHz.

[Baudry et al 1994] Baudry A., Lucas R., Guilloteau S., 1995, *A&A* **293**, 594
Accurate position of SiO masers.

[Bremer 1994] Bremer, M. 1994, IRAM internal report
The Phase Project: First Results

[Bremer 1995] Bremer, M., 1995, IRAM report **238**
The Phase Project: Observations on Quasars

[Bremer et al 1995] Bremer, M., Guilloteau, S., Lucas, R. 1995, IRAM Newsletter **24** (Nov. 1995)
Atmospheric Phase Correction: from random phase down to 20 degree r.m.s. at 230 GHz.

[Born and Wolf 1975] Born, M., and Wolf, E., 1975 *Pergamon Press*
Principles of Optics, New Physical Optics Notebook

[Cernicharo 1988] Cernicharo, J., 1998, PhD Thesis Paris, and IRAM report 52
ATM: A Program to compute Theoretical Atmospheric Opacity for Frequencies below 1 THz

[Cernicharo & Pardo 1999] Private communication.

[Christiansen and Hogbom 1969] Christiansen, W.N., and Hogbom, J.A., 1969, *Cambridge University Press*
Radiotelescopes

[Clark 1980] Clark, B.G., 1980, *A&A*, **89**, 377
An efficient implementation of the algorithm "CLEAN"

[Cooper 1970] Cooper, B.F.C., 1970, *Aust. J. Phys.* **23**, 521

[Cornwell 1987] Cornwell, T.J., 1987, MMA Memo. 42

[Cornwell 1988] Cornwell, T.J., 1988, *A&A*, **202**, 316
Radio interferometric imaging of very large objects.

[Cornwell 1989] Cornwell, T.J., 1989
in *Synthesis Imaging in Radio Astronomy*,
Perley R.A., Schwab F.R., Bridle A.H. (eds),
ASP Conferences Series Volume 6, p. 277

[Cornwell 1993] Cornwell, T.J., 1993
in *Astronomy with Millimeter and Sub-millimeter Wave Interferometry*,
Ishiguro M., Welch W.J. (eds),
ASP Conference Series Volume 59, p. 96

[Cotton & Schwab 1984] Schwab, F.R., 1984, *Astron. J.* **89**, 1076
Relaxing the isoplanetism assumption in self-calibration: applications to low frequency radio interferometry. (see top of p.1078)

[D'Addario 1989] D'Addario, L., 1989, in
Synthesis Imaging in Radio Astronomy,
Perley R.A., Schwab F.R., Bridle A.H. (eds),
ASP Conferences Series Volume 6, p. 59

[Downes & Altenhoff 1989] Downes, D., Altenhoff, W.J., 1989
Anomalous refraction at radio wavelengths
URSI/IAU Colloquium on Radio Astronomical Seeing, Beijing (China)
J.E.Baldwin, Wang Shouguan, eds., International Academic Publisher & Pergamon Press, p. 31

[Ekers & Rots 1979] Ekers, R.D., Rots, A.H., 1979, in
Images formation from coherence function in astronomy, C. van Schooneveld (ed.),
Proc. IAU Symp. 49, p. 61

[Goldsmith 1988] Goldsmith P.F., 1988, *IEEE press*
Instrumentation and Technique for Radio Astronomy

[Greve & Hooghoudt 1981] Greve A., Hooghoudt, B. 1981, *A&A* **93**, 76
Quality evaluation of radio reflector surface.

[Greve et al 1998] Greve A., Kramer, C., & Wild, W., 1998, *A&A Supp* **133**, 271
The beam pattern of the IRAM 30–m telescope (a reflector with several surface error distributions).

[Gueth 1997] Gueth, F., 1997, PhD thesis, University of Grenoble

[Gundlach 1989] Gundlach, K.H., 1989 (NATO ASI series, F 59, p. 259)

[Guilloteau et al 1993] Guilloteau, S., Dutrey, A., Marten, A., & Gautier, D., 1993, *A&A* **279**, 661
CO in the troposphere of Neptune: Detection of the J=1-0 line in absorption

[Guilloteau et al 1999] Guilloteau, S., Dutrey, A., Simon, M., 1999, *A&A* **348**, 570.
GG Tauri: the ring world.

[Hogbom 1974] Högbom, J.A., 1974, *A&A Supp.*, **15**, 417.
Aperture synthesis with a non regular distribution of interferometer baselines.

[Hagen et al 1973] Hagen, J.B., Farley, D.T., 1973, *Radio Science* **8**, 775

[Hill & Cliffort 1981] Hill, R.J., and Cliffort, S.F., 1981, *Radio Science* **16** No. 1, 77

[Karpov 1999] Karpov A., 1999, C.R.Acad.Sci.Paris, t.327, IIb, p.539

[Kraus 1982] Kraus, J.D., 1982 *McGraw Hill*
Radio Astronomy

[Kolmogorov 1941] Kolmogorov, A.N. (1941), in
Dissipation of energy in a locally isotropic turbulence, Doklady Akad. Nauk SSSR, **32**, 141
(English translation in: American Mathematical Society Translations 1958, Series 2, Vol 8, p. 87, Providence R.I)

[Lannes et al 1994] Lannes, A., Anterrieu, E., and Bouyoucef, K., *J. mod. Optics*, 1994, **41**, 1537-1574.
"Fourier Interpolation and Reconstruction via Shannon-type Techniques. Part I: Regularization principle."

[Lannes et al 1996] Lannes, A., Anterrieu, E., and Bouyoucef, K., *J. mod. Optics*, 1996, **43**, 105-138.
"Fourier Interpolation and Reconstruction via Shannon-type Techniques. Part II: Technical developments and applications"

[Lannes et al 1997] Lannes, A., Anterrieu, E., and Maréchal, P., *A&A Supp.*, 1997, **123**, 183-198.
CLEAN and WIPE

[Lay 1997] Lay, O., 1997, *A&A Supp* **122**, 547
Phase calibration and water vapor radiometry for millimeter wave arrays.

[Lo and Lee] Lo, S.W., & Lee
Antenna Handbook theory, applications and design

[Love 1978] Love, A.W., 1978, *IEEE press*
Reflector Antennas

[Lucas 1995] Lucas, R. 1995, IRAM internal report
Practical Implementation of phase correction.

[Mauersberger et al 1989] Mauersberger, R. et al 1989, *A&A Suppl* **79**, 217
Line calibrators at λ 1.3, 2 and 3 mm

[Martin-Pintado et al 1988] Martin-Pintado, J., et al 1988 *A&A* **197**, L15-L18.
Radiocontinuum and recombination lines towards CRL618 - Evidence for an ionized stellar wind ?

[Minnet and Thomas 1968] Minnet & Thomas 1968, *Proc. IEEE* **115,** 1419.

[Queney 1974] Elements de Météorologie, P. Queney, 1974,
Collection de l'Ecole Nationale de Techniques Avancées, Masson & Cie Pub. Paris

[Reynolds et al, 1989] Reynolds, G.O., DeVelis, J.B., Parrent, G.B.Jr., & Thompson, B.J., 1989, *SPIE Optical Engineering Press*
Physical optics notebook: Tutorial in Fourier optics.

[Ruze 1996] Ruze 1966, *Proc. IEEE* **54**, 633

[Sault et al 1996] Sault R.J., Staveley-Smith L., Brouw W.N., 1996, *A&A Supp.*, **120**, 375
An approach to interferometric mosaicing

[Schwarz 1978] Schwarz U.J., 1978, *A&A*, **65**, 345
Mathematical-statistical description of the iterative beam removing technique (method CLEAN).

[Shannon 1949] Shannon, C.E., 1949, *Proc. IRE*, **37**, 10-21

[Sovers et al 1998] Sovers O.J., Fanselow J.L., Jacobs C.S., in *Reviews of Modern Physics*, Vol. **70**, No. 4, October 1998, p.1393.

[Steer et al 1984] Steer D.G., Dewdney P.E., Ito M.R., 1984, *A&A*, **137**, 159
Enhancement to the deconvolution algorithm "CLEAN".

[Tatarski 1961] Tatarski V.I. 1961
Wave Propagation in a turbulent Medium, McGraw-Hill

[Tatarski 1971] Tatarski V.I. 1971
The Effects of the turbulent Atmosphere on Wave Propagation
(translated from Russian by the Israel Program for Scientific Translations Ltd, ISBN 0 7065 0680 4)
reproduced by National Technical Information Service, U.S. Department of Commerce, Springfield, Va. 22151

[Thayer 1974] Thayer G.D. 19974, *Radio Science* **9** No 10, p. 803

[Thompson et al 1986] Thompson, A.R., Moran, J.M., Swenson, G.W., 1986
Interferometry and Synthesis in Radio Astronomy (Wiley : New York)

[Townes & Schawlow 1975] C.H. Townes, A.L. Schawlow, 1975
Microwave Spectroscopy, Dover Pub. New York

[Van Vleck 1966] Van Vleck, J.H., Middleton, D., 1966, *Proc. IEEE* **54**, 2

[Waters 1976] Absorption and Emission by atmospheric gases, J.W.Waters 1976,
in *Methods of experimental physics*, **12**, ed. M.L.Meeks p. 147 Academic Press, New York

[Wakker & Schwartz 1988] Wakker B., Schwarz U.J., 1988, *A&A*, **200**, 312
The multi-resolution CLEAN and its application to the short-spacing problem in interferometry.