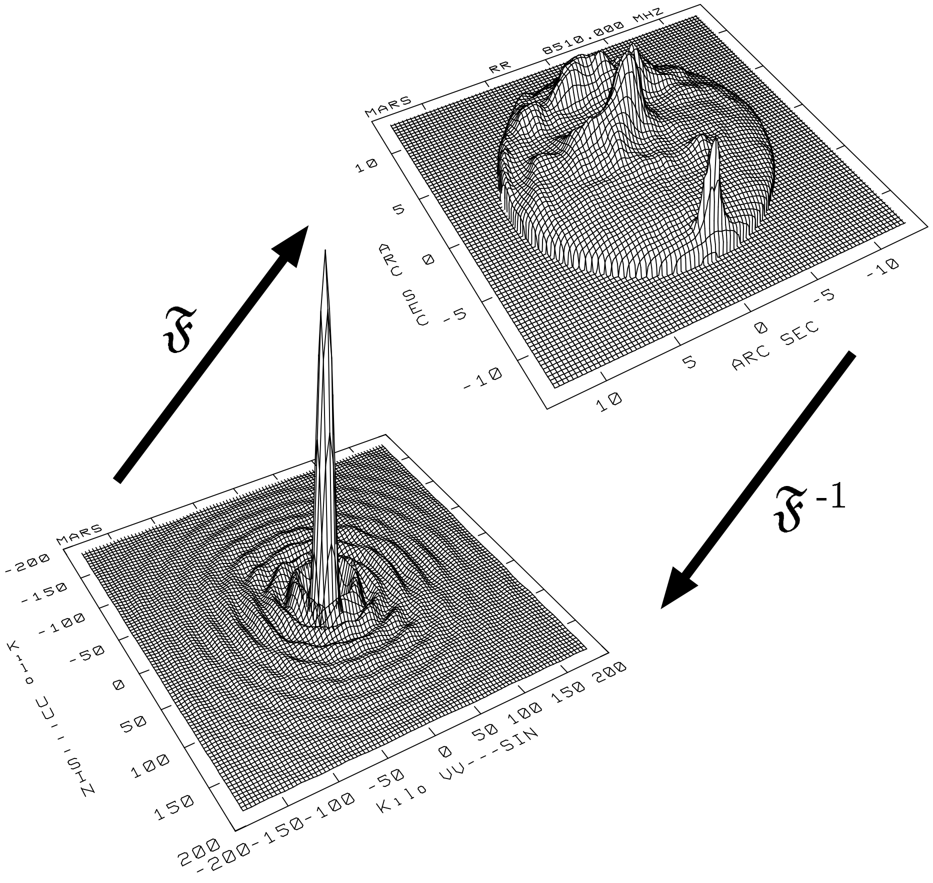
Synthesis Imaging in Radio Astronomy II



A Collection of Lectures from the Sixth NRAO/NMIMT Synthesis Imaging Summer School. Held in Socorro NM 1998 June 17–23.

Edited by

G. B. Taylor, C. L. Carilli, and R. A. Perley

Cover Illustration: Image of 3.5cm radar reflectivity of Mars made using Goldstone as the transmitting telescope and the VLA as the receiving telescope, along with its Fourier transform (visibility function). Bright reflections from south polar ice and the Tharsis and Olympus Mons volcanoes and associated lava flows are apparent as positive relief on the image. Data courtesy of B.J.Butler, D.O.Muhleman, and M.A.Slade.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

Contents

Contents

Lecture 15. Error Recognition *by R. D. Ekers*

Lecture 16. Non-Imaging Data Analysis *by T. J. Pearson*

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

Preface

The Summer School on *Synthesis Imaging* held in Socorro, New Mexico from June 17 to June 23, 1998 was the sixth in a series held approximately every three years. The lectures were given by the staff of the National Radio Astronomy Observatory and invited speakers from the University of Illinois, the California Institute of Technology, the Harvard University, the University of Maryland, the New Mexico Institute of Mining and Technology, and the Australia Telescope National Facility. This Volume contains the edited texts of lectures from the series, and succeeds the previous collection from the Third Synthesis Imaging Summer School published in 1989. It is intended for serious students of synthesis imaging and image processing.

Purpose of the course

The NRAO operates two of the world's most powerful synthesis radio telescopes, the Very Large Array (VLA), located about 50 miles from Socorro on the Plains of San Augustin, and the Very Long Baseline Array (VLBA), a synthesis telescope whose elements are distributed from St. Croix to Hawaii. Both telescopes are operated from Socorro. The major goal of this course, like that of its predecessors beginning in 1982 was to inform potential users of these two synthesis telescopes about the principles of these instruments' operation, about subtleties of data acquisition, calibration and processing with such instruments, and about techniques for obtaining the best results from them.

As such, the course is aimed first at radio astronomers who need synthesis techniques and instruments for their research. It is hard for graduate students to get practical experience of synthesis imaging outside the national facilities, and many undergraduate physics and engineering curricula treat only the barest outline of Fourier methods and coherence. Many of the exciting developments in data processing that have made synthesis telescopes so powerful also receive only terse treatment in the general literature or in textbooks. We therefore have set out to discuss the subject as fully as necessary for those who wish to use the NRAO's radio synthesis instruments for their own research. Our goal is to discuss the subject in enough detail that the student can appreciate both the strengths and limitations of the synthesis technique, and so begin to evaluate how much, or how little, credence to give individual synthesis images. To exploit an instrument fully for frontier research, the user must understand it thoroughly enough to distinguish unexpected instrumental errors from unexpected discoveries about the cosmos. A firm understanding of synthesis instruments involves understanding their operating principles (and the assumptions that underlie them), their hardware and the algorithms and software used in the data reduction. All these topics are covered here.

Although synthesis imaging is a specialized skill even within radio astronomy, it is grounded in mathematical and physical principles that have applications in other fields. This course may therefore also be relevant beyond the traditional community of radio astronomers, for example to anyone wishing to interpret images made by synthesis telescopes. It may also interest researchers who use Fourier methods or deconvolution techniques for imaging in physics, medicine, remote sensing, seismology or radar.

We have repeated the course at three-year intervals, both because the subject matter is still evolving and maturing and because a new generation of graduate students passes across the national scene every three or four years. Our three-year cycle matches this generational change in the students and also provides a manageable number of new developments to introduce and old topics to reconsider each time.

Subject matter

In this 1998 version of the course we again divided the material into two main parts, which were given in successive weeks, with a "hands on" demonstration at the Array Operations Center during the intervening weekend.

The first segment of the course contains sixteen lectures that describe the fundamentals of synthesis imaging and which could be read as a stand-alone course by the beginning student. The second segment consists of seventeen lectures on more advanced and specialized topics. Unlike previous publications in this series that gave only cursory treatment of subjects specific to VLBI, in this Volume Lectures 22 (VLBI), 23 (Astrometry and Geodesy), 24 (Spectral Line VLBI), 25 (VLBI Polarimetry) and 26 (Space VLBI), cover many fundamental topics specifically related to VLBI. Furthermore an attempt was made to make the lectures in this 1998 version more generally applicable to both connected element and Very Long Baseline interferometry. Examples from the VLA, however, still dominate the first segment of the book.

The lectures do not appear here exactly as they were given. The lecturers reviewed comments from other NRAO staff and from the editors before finalizing their texts. Difficult points have been explained in greater detail, and additional material that could not be covered within the time constraints of the live lectures has been added. The editors have also standardized notation and have rearranged material where we felt that this would add coherence to the course as a whole. For the reader's convenience, we also include a comprehensive index at the end of the book.

NRAO lore

There are references to NRAO internal publications and to NRAO software throughout this course. These references will be important further reading for VLA and VLBA users. Copies of memoranda in the VLA and VLBA technical and scientific series are available on request from Gayle Rhodes, NRAO, P.O. Box O, Socorro, NM 87801, U.S.A. Information about the NRAO Astronomical Image Processing System (AIPS) software—and the software itself—can be obtained on request to the AIPS Group, NRAO, Edgemont Road, Charlottesville, VA 22903-2475, U.S.A. There is also an ever-increasing body of information on all aspects of proposing and observing with the VLA and VLBA available on the World Wide Web starting from `http://www.nrao.edu`.

Acknowledgments

The School was co-sponsored by the New Mexico Institute of Mining and Technology, and by Associated Universities, Inc. We are grateful to those organizations for their sponsorships.

We are indebted to all the school Lecturers who submitted manuscripts for this Volume. We especially thank authors for their efforts in producing high-quality manuscripts and in achieving bibliographic completeness and correctness. We are also indebted to the editors of the 1989 volume for producing an excellent work that guided our efforts. In particular we thank former editor Fred Schwab for comments on a preliminary draft of this book. We acknowledge the use of the on-line bibliographic Abstract Service of the Astrophysics Data System at Harvard University. We also owe thanks to L. Appel, J. Chavez, T. Romero, and S. Lagoyda for their assistance preparing this Volume and organizing the school itself.

G. B. TAYLOR
C. L. CARILLI
R. A. PERLEY

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

Dedication

This book is dedicated to the memory of Daniel S. Briggs. Dan was killed in a skydiving accident near Chicago just a few weeks after the summer school. Dan was an enthusiastic teacher and researcher who will be sorely missed. His 'Robust' weighting scheme has become the standard for synthesis imaging. At the summer school he gave an excellent lecture on imaging and deconvolution (see Lectures 7 and 8 of this volume contributed by Dan and collaborators). He entered thoroughly into the spirit of the summer school – attending all the lectures and interacting with many of the participants. Dan also influenced the style of this book by making many helpful suggestions. We are in his debt.

The above picture of Dan was taken at the school by Nick Dudish.

G. B. Taylor
C. L. Carilli
R. A. Perley

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

1. Coherence in Radio Astronomy

B. G. Clark
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract.
In this lecture the main principles of synthesis imaging are derived.

1. Introduction

It is appropriate for this specialized summer school to start with a survey of the derivation of the main principles of synthesis imaging, paying particular attention to the assumptions that go into them. This is because a substantial number of the lectures to follow will discuss the problems which arise when these assumptions are violated under the conditions of the observation the astronomer wants to make. At the same time, I will cast this introduction into the terminology of modern optics, in an attempt to stay abreast of current fashions in physics.

The fundamental reference for the basics of modern optics is the excellent textbook *Principles of Optics*, by Born and Wolf; their Chapter X is especially relevant to this summer school; another excellent reference on physical optics is *Statistical Optics* by Goodman. An excellent discussion of synthesis imaging, employing this modern terminology, is given by J. L. Yen (1985) in Chapter 5 of *Array Signal Processing* (S. Haykin, Ed.). A broader view of radio telescopes, again from a viewpoint of Fourier optics, but taking a somewhat historical perspective, is given in *Radiotelescopes* by Christiansen & Högbom (1985, Second Edition); their Chapter 7 discusses synthesis methods. The alternate viewpoint on radio interferometers, from the perspective of the electrical engineers who originally developed them, is explicated in Swenson & Mathur (1968).

2. Form of the Observed Electric Field

I will start with the most general formulation of the subject, and, one by one, introduce the simplifying assumptions until reaching the simple, elegant theoretical basis that is, after all, sufficient for much of radio interferometry. In the most general case, an astrophysical phenomenon occurs at location $\mathbf{R}$ (the boldface symbols indicate vectors, in this case a position vector). This phenomenon causes a time-variable electric field, which I will denote as $\mathbf{E}(\mathbf{R}, t)$. Then, Maxwell's equations say that an electromagnetic wave will propagate away from that point and will eventually arrive at a point where an astronomer may conveniently observe it, say the point $\mathbf{r}$. See Figure 1–1.

It is inconvenient for a number of reasons to deal with general time-variable electric fields. If we have a finite time interval of a varying field, we may express the magnitude of the field as the real part of the sum of the Fourier series in which the only time-varying functions are simple exponentials. Because of the linearity of Maxwell's equations (in the cases of astrophysical interest, anyway) we may deal with the coefficients of this Fourier series, rather than with the general time-varying field. The coefficients of this Fourier series, which I will

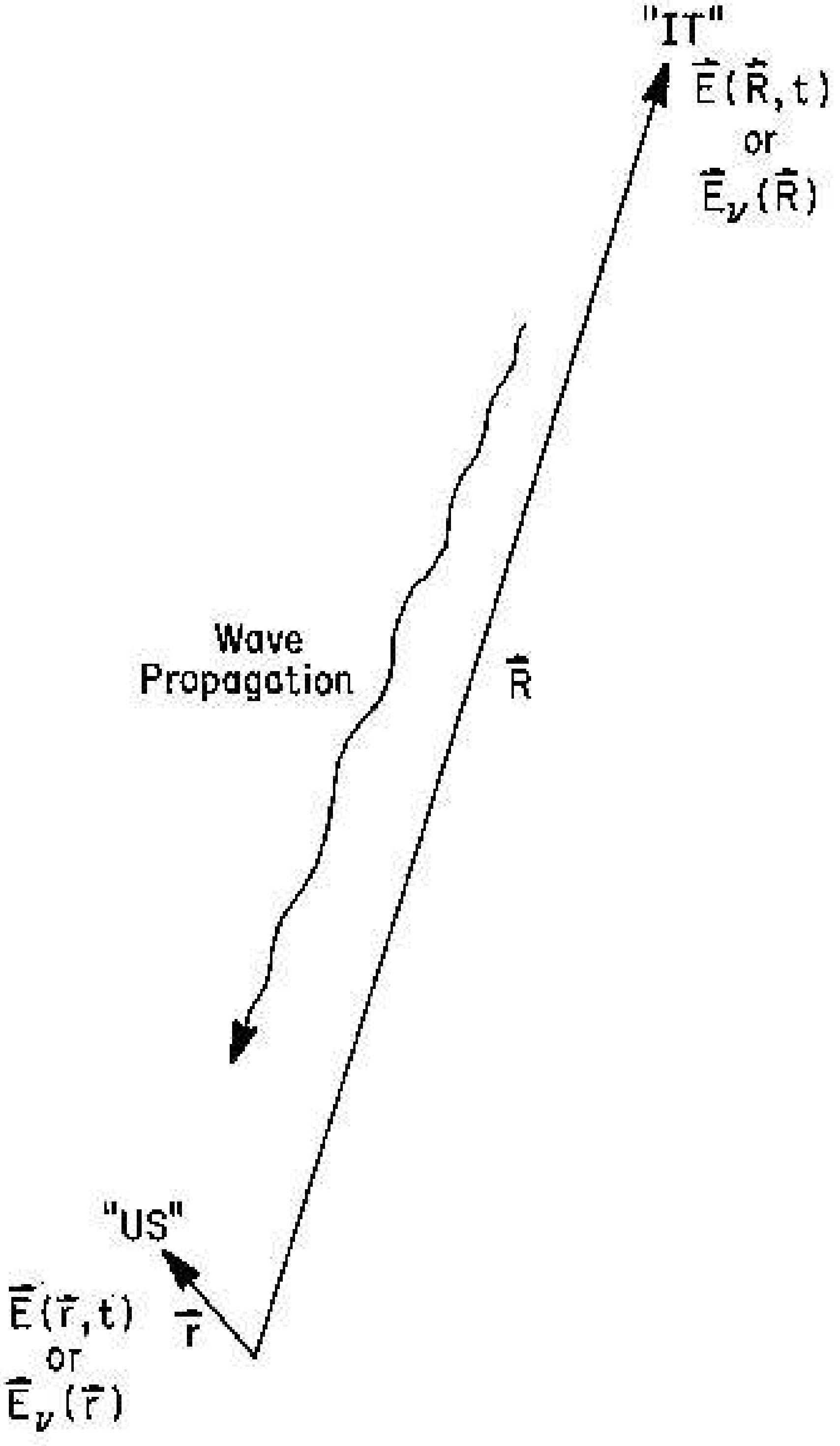


Figure 1–1. An astrophysically interesting phenomenon "IT", located at **R**, causes an electromagnetic wave $\mathbf{E}_\nu(\mathbf{r})$ to propagate through space, where it can be detected at **r** by "US".

call $\mathbf{E}_\nu(\mathbf{R})$, are called the *quasi-monochromatic components* of the electric field $\mathbf{E}(\mathbf{R}, t)$. Note that the fields $\mathbf{E}_\nu(\mathbf{R})$ are complex quantities, and it is useful to think of them as such at all times. It leads to a more elegant formulation of the theories to consider this complex nature to be physical reality rather than a mathematical convenience.

In what follows, I consider only a single quasi-monochromatic component, realizing that the total response is the sum of the responses to all the components. In fact, the response of an instrument can be made arbitrarily close to that of a single quasi-monochromatic component, by inserting filters in the early, linear, parts of the instrument.

The linearity property of Maxwell's equations allows us to superpose the fields produced at a test location by the various source points,

$$\mathbf{E}_\nu(\mathbf{r}) = \iiint P_\nu(\mathbf{R}, \mathbf{r}) \mathbf{E}_\nu(\mathbf{R}) \, dx \, dy \, dz \, . \tag{1–1}$$

The integral is to be taken over all of space. The function $P_\nu(\mathbf{R}, \mathbf{r})$ is called the *propagator*, and describes how the electric field at $\mathbf{R}$ influences the electric field at $\mathbf{r}$.

At this point, I begin to introduce the simplifying assumptions. The first assumption may be considered to be merely pedagogical, in the sense that it is not really needed at all, and is made only to avoid complicating the equations to the point that their physical meaning is obscured. For the moment, I shall ignore the fact that electromagnetic radiation is a vector phenomenon, and treat it as if it were simply a scalar field, measured at any point by a scalar quantity E. That is to say, I shall ignore, for the moment, all polarization phenomena. This enables the multiplication in Equation 1–1 to be regarded as ordinary scalar multiplication, and the propagator P to be an ordinary scalar function (not a tensor function as a complete derivation would have it).

The second simplifying assumption is that the sources of interest to astronomers are a long way away. The practical implication of this is that we have to give up all hope of describing the structure of the emitting regions in the third dimension, in depth. All we may measure is the "surface brightness" of the emitting phenomenon. One convenient way of expressing this assumption is to replace the field strength E at the source with the field strength at a convenient point distant from both us and from any source of radiation. We may conceive of a "celestial sphere", a very large sphere of radius $|\mathbf{R}|$, within which there is no additional radiation, and all that we may learn about the distribution of the source of the fields is the distribution of the electric field on the surface of this sphere, which I will call $\mathcal{E}_\nu(\mathbf{R})$. See Figure 1–2.

The third simplifying assumption is that the space within the "celestial sphere" is empty. For this case, Huygens' Principle tells us that the propagator takes a particularly simple form, and we can write

$$E_\nu(\mathbf{r}) = \int \mathcal{E}_\nu(\mathbf{R}) \frac{e^{2\pi i \nu |\mathbf{R}-\mathbf{r}|/c}}{|\mathbf{R}-\mathbf{r}|} \, dS \, . \tag{1–2}$$

Here dS is the element of surface area on the celestial sphere.

Equation 1–2 is the general form of the quasi-monochromatic component of the electric field at frequency ν due to all sources of cosmic electromagnetic

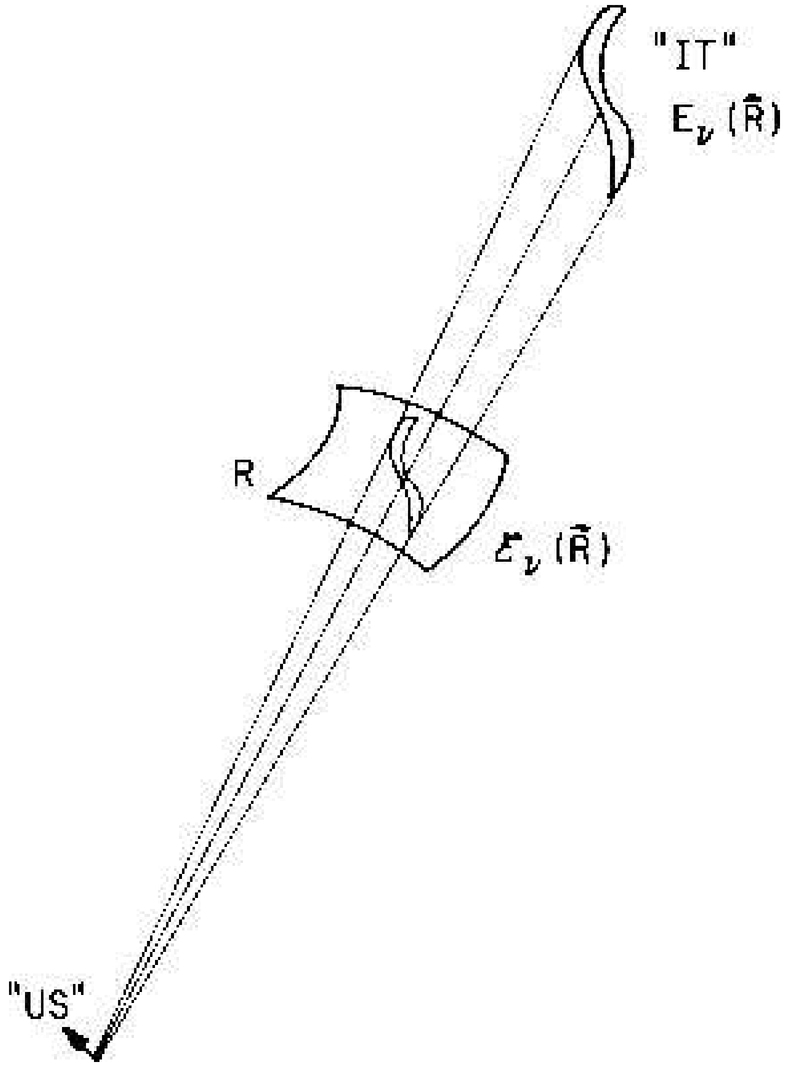


Figure 1–2. The passage of radiation from a distant source through an imaginary sphere at radius $|\mathbf{R}|$ defines the electric field distribution $\mathcal{E}_\nu(\mathbf{R})$ on this surface. For the purposes of synthesis imaging, all astronomical sources may be considered to lie on this sphere, provided only that their real distances greatly exceed B^2/λ, where $B = |\mathbf{r_1} - \mathbf{r_2}|$ is the baseline length.

radiation. This is all we have; we can measure only the properties of this field $E_\nu(\mathbf{r})$ to tell us about the nature of things at large in the universe.

3. Spatial Coherence Function of the Field

Among the properties of $E_\nu(\mathbf{r})$ is the correlation of the field at two different locations. The correlation of the field at points $\mathbf{r_1}$ and $\mathbf{r_2}$ is defined as the expectation of a product, namely $\mathsf{V}_\nu(\mathbf{r_1}, \mathbf{r_2}) = \langle \mathbf{E}_\nu(\mathbf{r_1}) \mathbf{E}_\nu^*(\mathbf{r_2}) \rangle$. The raised asterisk indicates the complex conjugate. We can then use Equation 1–2 to substitute for $E_\nu(\mathbf{r})$, writing the product of the integrals as a double integral over two separate surface element dummy variables:

$$\mathsf{V}_\nu(\mathbf{r}_1, \mathbf{r}_2) = \left\langle \iint \mathcal{E}_\nu(\mathbf{R}_1)\mathcal{E}^*_\nu(\mathbf{R}_2) \frac{e^{2\pi i\nu|\mathbf{R}_1-\mathbf{r}_1|/c}}{|\mathbf{R}_1 - \mathbf{r}_1|} \frac{e^{-2\pi i\nu|\mathbf{R}_2-\mathbf{r}_2|/c}}{|\mathbf{R}_2 - \mathbf{r}_2|} \, dS_1 \, dS_2 \right\rangle . \tag{1–3}$$

The fourth simplifying assumption is that the radiation from astronomical objects is not spatially coherent; i.e., that $\langle \mathcal{E}_\nu(\mathbf{R}_1)\mathcal{E}^*_\nu(\mathbf{R}_2)\rangle$ is zero for $\mathbf{R}_1 \neq \mathbf{R}_2$. Exchanging the expectation and the integrals in Equation 1–3 then gives:

$$\mathsf{V}_\nu(\mathbf{r}_1, \mathbf{r}_2) = \int \langle |\mathcal{E}_\nu(\mathbf{R})|^2\rangle |\mathbf{R}|^2 \frac{e^{2\pi i\nu|\mathbf{R}-\mathbf{r}_1|/c}}{|\mathbf{R} - \mathbf{r}_1|} \frac{e^{-2\pi i\nu|\mathbf{R}-\mathbf{r}_2|/c}}{|\mathbf{R} - \mathbf{r}_2|} \, dS \, . \tag{1–4}$$

Now write $\mathbf{s}$ for the unit vector $\mathbf{R}/|\mathbf{R}|$ and $I_\nu(\mathbf{s})$ for the observed *intensity* $|\mathbf{R}|^2\langle|\mathcal{E}_\nu(\mathbf{s})|^2\rangle$. The second assumption (the great distance to the sources and to the celestial sphere) can then be used again to neglect the small terms of order $|\mathbf{r}/\mathbf{R}|$, and to replace the surface element dS on the celestial sphere by $|\mathbf{R}|^2 d\Omega$, so that Equation 1–4 becomes:

$$\mathsf{V}_\nu(\mathbf{r}_1, \mathbf{r}_2) \approx \int I_\nu(\mathbf{s}) e^{-2\pi i\nu \mathbf{s}\cdot(\mathbf{r}_1-\mathbf{r}_2)/c} \, d\Omega \, . \tag{1–5}$$

Observe that Equation 1–5 depends only on the separation vector $\mathbf{r}_1 - \mathbf{r}_2$ of the two points, not on their absolute locations $\mathbf{r}_1$ and $\mathbf{r}_2$. Therefore, we can find out all we can learn about the correlation properties of the radiation field by holding one observation point fixed and moving the second around; we do not have to measure at all possible pairs of points. This function V_ν of a single (vector) separation $\mathbf{r}_1 - \mathbf{r}_2$ is called the *spatial coherence function*, or the *spatial autocorrelation function*, of the field $E_\nu(\mathbf{r})$. It is all we have to measure.

An interferometer is a device for measuring this spatial coherence function.

4. The Basic Fourier Inversions of Synthesis Imaging

A second interesting point about Equation 1–5 is that the equation is, within reasonable, well-defined limits, invertible. The intensity distribution of the radiation as a function of direction $\mathbf{s}$ can therefore be deduced in certain cases by measuring the spatial coherence function V as a function of $\mathbf{r}_1 - \mathbf{r}_2$ and performing the inversion.

There are two special cases of a great deal of interest. In fact, it is usually argued that any actual case is so close to one of these two special cases that the invertibility properties (although not necessarily the effort required to perform the inversion) must be essentially similar. Since there are two forms of interest, there are two alternate forms of our fifth (and final) simplifying assumption.

4.1. Measurements confined to a plane

First, we could choose to make our measurements only in a plane; that is, in some favored coordinate system, the vector spacing of the separation variable in the coherence function, conveniently measured in terms of the wavelength $\lambda = c/\nu$, is $\mathbf{r}_1 - \mathbf{r}_2 = \lambda(u, v, 0)$. In this same coordinate system, the components of the unit vector $\mathbf{s}$ are $(l, m, \sqrt{1 - l^2 - m^2})$. Inserting these, and explicitly showing the form, in this coordinate system, of the element of solid angle, we have

$$V_\nu(u, v, w \equiv 0) = \iint I_\nu(l, m) \frac{e^{-2\pi i(ul+vm)}}{\sqrt{1 - l^2 - m^2}} \, dl \, dm \,. \tag{1–6}$$

Equation 1–6 is, clearly, a Fourier transform relation between the spatial coherence function $V_\nu(u, v, w \equiv 0)$ (with separations expressed in wavelengths), and the modified intensity $I_\nu(l, m)/\sqrt{1 - l^2 - m^2}$ (with angles expressed as direction cosines). Now we are home free. Mathematicians have devoted decades of work to telling us when we can invert a Fourier transform, and how much information it requires.

4.2. All sources in a small region of sky

The alternate form of the fifth simplifying assumption is to consider the case where all of the radiation comes from only a small portion of the celestial sphere. That is, to take $\mathbf{s} = \mathbf{s}_0 + \sigma$, and neglect all terms in the squares of the components of σ. In particular, the statement that both $\mathbf{s}$ and $\mathbf{s}_0$ are unit vectors implies that

$$\begin{aligned} 1 = |\mathbf{s}| = \mathbf{s} \cdot \mathbf{s} &= \mathbf{s}_0 \cdot \mathbf{s}_0 \\ &= \mathbf{s}_0 \cdot \mathbf{s}_0 + 2\mathbf{s}_0 \cdot \sigma + \sigma \cdot \sigma \\ &\approx 1 + 2\mathbf{s}_0 \cdot \sigma \,, \end{aligned}$$

i.e., $\mathbf{s}_0$ and σ are perpendicular. If we again introduce a special coordinate system such that $\mathbf{s}_0 = (0, 0, 1)$, then we have a slightly different offspring of Equation 1–5:

$$V'_\nu(u, v, w) = e^{-2\pi i w} \iint I_\nu(l, m) e^{-2\pi i(ul+vm)} \, dl \, dm \,. \tag{1–7}$$

Here, the components of the vector $\mathbf{r}_1 - \mathbf{r}_2$ have been denoted by (u, v, w). It is customary to absorb the factor in front of the integral in Equation 1–7 into the left hand side, by considering the quantity $V_\nu(u, v, w) = e^{2\pi i w} V'_\nu(u, v, w)$, which we see is independent of w:

$$V_\nu(u, v) = \iint I_\nu(l, m) e^{-2\pi i(ul+vm)} \, dl \, dm \,. \tag{1–8}$$

$V_\nu(u, v)$ is the coherence function relative to the direction $\mathbf{s}_0$, which is called the *phase tracking center.*

Since Equation 1–8 is a Fourier transform, we have in particular, the direct inversion

$$I_\nu(l, m) = \iint V_\nu(u, v) e^{2\pi i(ul+vm)} \, du \, dv \,. \tag{1–9}$$

The relationship between the two different forms of the assumption used here and in Section 4.1 can be seen from the symmetric role played in Equation 1–5 by the two vectors $\mathbf{s}$ and $\mathbf{r}_1 - \mathbf{r}_2$: the form used in Section 4.1 results from saying that the *vectors* $\mathbf{r}_1 - \mathbf{r}_2$ lie in a plane; the form used here results from saying that the *endpoints* of the vectors $\mathbf{s}$ lie in a plane.

4.3. Effect of discrete sampling

In practice the spatial coherence function V is not known everywhere but is sampled at particular places on the u-v plane. The sampling can be described by a *sampling function* $S(u, v)$, which is zero where no data have been taken. One can then calculate a function

$$I_\nu^D(l, m) = \iint V_\nu(u, v) S(u, v) e^{2\pi i(ul+vm)}\, du\, dv\,. \tag{1–10}$$

Radio astronomers often refer to $I_\nu^D(l, m)$ as the *dirty image*; its relation to the desired intensity distribution $I_\nu(l, m)$ is (using the convolution theorem for Fourier transforms):

$$I_\nu^D = I_\nu * B\,, \tag{1–11}$$

where the in-line asterisk denotes convolution and

$$B(l, m) = \iint S(u, v) e^{2\pi i(ul+vm)}\, du\, dv \tag{1–12}$$

is the *synthesized beam* or *point spread function* corresponding to the sampling function $S(u, v)$. Equation 1–11 says that I^D is the true intensity distribution I convolved with the synthesized beam B. Lecture 8 discusses methods for undoing this convolution.

4.4. Effect of the element reception pattern

An additional minor alteration must be made to the above for convenience in practical calculation. In practice, the interferometer elements are not point probes which sense the voltage at that point, but are elements of finite size, which have some sensitivity to the direction of arrival of the radio radiation. That is, there is an additional factor within the integral of Equation 1–2, and hence of Equations 1–4, 1–5, 1–6, 1–7 and 1–8, of $\mathcal{A}_\nu(\mathbf{s})$ (the *primary beam* or *normalized reception pattern* of the interferometer elements) describing this sensitivity as a function of direction. For explicitness, Equation 1–8 is rewritten below, with this factor included:

$$V_\nu(u, v) = \iint \mathcal{A}_\nu(l, m) I_\nu(l, m) e^{-2\pi i(ul+vm)}\, dl\, dm\,. \tag{1–13}$$

The $V_\nu(u, v)$ so defined is normally termed the *complex visibility* relative to the chosen phase tracking center.

It is clear that dealing with the element directivity $\mathcal{A}_\nu$ should be postponed to the final step of deriving the sky intensity, and that then it should simply divide the derived intensities (if all elements have the same reception pattern). This division will, however, not only produce a better estimate of the actual intensities in this direction, but will also increase the errors (of all types) in directions far from the center of the element primary beam, where one is dividing by small numbers.

Although the factor $\mathcal{A}_\nu$ looks like merely a nuisance, it is actually the reason that the second form of the final assumption (used in Section 4.2) is so acceptable in many practical cases—$\mathcal{A}_\nu(\mathbf{s})$ falls rapidly to zero except in the vicinity of some $\mathbf{s}_0$, the pointing center for the array elements.

5. Extensions to the Basic Theory

Two simple extensions to this basic theory are worth mentioning at this point.

5.1. Spectroscopy

First, consider the case of observing a spectral line. Here the appearance of the sky may change quite rapidly as a function of frequency, and one would like to make synthesis images at a large number of closely spaced frequencies. Clearly, one can do this by inserting narrowband filters into the early, linear, parts of the interferometer, and simply repeat the processing for each frequency, either seriatim or simultaneously. However, there is a technically more attractive approach. With current technology, it is attractive to implement the latter portions of the interferometer in digital hardware. In this technology, it is quite inexpensive to add additional multipliers to calculate the correlation as a function of lag. Admitting a range of quasi-monochromatic waves to the interferometer, we can write an expression for the correlation as a function of lag, noting that for each quasi-monochromatic wave, a lag is equivalent to a phase shift, i.e., a multiplication by a complex exponential

$$V(u, v, \tau) = \int V(u, v, \nu) e^{2\pi i \tau \nu} \, d\nu \, . \tag{1–14}$$

The above is clearly a Fourier transform, with complementary variables ν and τ, and can be inverted to extract the desired $V(u, v, \nu)$. Since, in this digital technology, one is dealing with sampled data, I give the sampled form of the inversion below:

$$V(u, v, j\Delta\nu) = \sum_k V(u, v, k\Delta\tau) e^{-2\pi i j k \Delta\nu \Delta\tau} \, . \tag{1–15}$$

The fact that we are dealing with sampled data is of some interest, and we should stop and inquire about how the Fourier sampling theorem is to be applied. Examining the above, in its full complex form, we see that the replication interval is $1/\Delta\tau$ in frequency, so that the band of frequencies must be limited, before sampling, to a total bandwidth of less than this, to avoid loss of information in the sampling process.

This is different from the statement one usually encounters, in which a prefiltering to $1/2\Delta\tau$ is required to preserve the information in the sampling process for a signal (actually it is usually stated, equivalently, as requiring a sampling interval of $1/2B$, where B is the prefilter bandwidth). This factor of two difference is due to the complex nature of the quantities we have been dealing with—the $V(u, v, \nu)$ are complex numbers, calculated by a complex multiplication of the complex field quantities. Complex multipliers and complex samplers require at least twice as many electronic components as devices that produce a real number, and the resulting doubling of the hardware permits us to sample a factor of two less often.

However, one can also develop this theory from the conventional viewpoint of dealing with real numbers only. Here the $2B$ sample rate is required, and maintains all the information in the signals. We can exchange this faster sampling rate for the double hardware needed to produce the complex version of the signals. The real parts of the various $V(u, v, \nu)$ are derived from the part of the correlation function that is even about $\tau = 0$, and the odd part supplies the imaginary part of $V(u, v, \nu)$.

Finally, if one derives the spectrum in this manner, one can, clearly, convert back to the single continuum channel at zero lag simply by summing the derived frequency-dependent V. This process clearly results in a complex number, even though each measurement was only a real number. The process of transforming a real function into a complex one by Fourier transforming and then transforming back on half the interval is called a Hilbert transform, and is an alternate method to implementing complex correlators.

5.2. Polarimetry

Now, in a final remark, let me look back to Section 2, to the first simplifying assumption, that of a scalar field. Actually, the electromagnetic field is a vector phenomenon, and the polarization properties carry interesting physical information. For the case of noise emission, one must be a bit careful about the definition of polarization. A monochromatic wave is always completely polarized, in some particular elliptical polarization, in that a single number describes the variation of the fields everywhere. For electromagnetic noise, polarization is defined by a correlation process. One picks two orthogonal polarizations and analyzes the radiation of the quasi-monochromatic waves into the components in these two polarizations. Then the polarization of the quasi-monochromatic wave is described by the 2 × 2 matrix of correlations between these two resolutions into orthogonal components. For instance, if we pick right and left circular polarization as the two orthogonal modes, then the matrix

$$\begin{pmatrix} RR^* & RL^* \\ LR^* & LL^* \end{pmatrix} \tag{1–16}$$

describes the polarization. This can be related to more familiar descriptions of polarization. For instance, the *Stokes parameters* have the intensity I, two linear polarization parameters Q and U, and a circular polarization parameter V related to the above numbers in simple (and more or less obvious) linear combinations:

$$\begin{pmatrix} I + V & Q + iU \\ Q - iU & I - V \end{pmatrix} \tag{1–17}$$

The complex correlation functions on the celestial sphere are preserved in the spatial coherence functions that interferometers measure. That is, one can derive, for instance, the distribution of $\langle RR^* \rangle$ on the sky by measuring the coherence function of R on the ground, and so forth for the other components of the matrix in (1–16). Since the intensity is the quantity in which one is always interested, one usually forms the sum of the R and L coherence functions before transforming to the sky plane, which one can always do, since the relations are linear. One can choose to do the same with the other Stokes parameters, or one can calculate the transforms of the mutual coherence between R and L to find the distribution of $\langle RL^* \rangle$ on the sky, and later note that this is, in terms of the Stokes parameters, $Q + iU$.

References

Born, M. & Wolf, E. 1980, *Principles of Optics*, Sixth (Corrected) Edition, Pergamon Press (Oxford, England).

Christiansen, W. N. & Högbom, J. A. 1985, *Radiotelescopes*, Second Edition, Cambridge University Press (Cambridge, England).

Goodman, J. W. 1985, *Statistical Optics*, John Wiley & Sons, New York.

Swenson, G. W., Jr. & Mathur, N. C. 1968, *Proc. IEEE*, 56, 2114–2130.

Yen, J. L. 1985, in *Array Signal Processing*, S. Haykin, Ed., Prentice–Hall (Englewood Cliffs, NJ), pp. 293–350.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

2. Fundamentals of Radio Interferometry

A. Richard Thompson
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A.

Abstract.
The practical aspects of interferometry are reviewed, starting with a two element interferometer.

1. Introduction

In the first lecture, it was shown that images of a distant radio source can be made by measuring the mutual coherence function of the electric fields at pairs of points in a plane normal to the direction to the source. This process can be envisioned by considering a large flat area on the Earth's surface on which antennas are located, and an electronic system including correlators to measure the coherence of the received signals. Now suppose that the Earth does not rotate, and that the source under observation is at the zenith. We can measure the coherence as a function of the spacing between pairs of antennas, independent of the absolute location of the antennas in the antenna plane. Next suppose that the Earth remains fixed relative to the sky, but that the source under investigation is not at the zenith. A wavefront from the source meets the plane in a line that progresses across the plane, and thus does not reach all antennas simultaneously. As a result, we need to include delay elements in the electronic system to ensure that the signals received by different antennas from the same wavefront arrive at the correlators at the same time. The spacings in (u, v) coordinates are now the components of the baselines projected onto a plane normal to the direction of the source. Finally, consider the effect of the rotation of the Earth. It is necessary for the antenna pointing and the time delays to be continuously adjusted to follow the source across the sky. Further, although at any instant the baselines lie within a plane, this plane is carried through space by the Earth, and its position relative to the source is continuously changing. The non-coplanar distribution of the baselines which thus generally[1] occurs during an extended period of observation can complicate the inversion of the coherence data to obtain an image. On the other hand, the motion of the baseline vectors has the useful effect of reducing the number of antenna locations on the Earth required to measure the coherence over the necessary range of (u, v) coordinates.

In this lecture the effects described above will be examined in some detail, as part of a discussion of the practical implementation of interferometers and arrays. A number of the relationships derived in the treatment deriving from physical optics will emerge again as expressions for the response of an interferometer. Other discussions of the interferometer response can be found in Swenson & Mathur (1968), Fomalont (1973), Fomalont & Wright (1974), Meeks (1976), Christiansen & Högbom (1985), and Rohlfs (1986); a detailed and extensive review is given by Thompson, Moran & Swenson (1986).

[1] Except in the case of an East–West linear array, as will be explained.

2. Response of an Interferometer

Synthesis arrays, which produce images by Fourier synthesis from measurements of complex visibility, can be analyzed as ensembles of two-element interferometers. Many of the effects can therefore be understood from a discussion of the properties of a two-element instrument. A simplified block diagram of such an interferometer is shown in Figure 2–1. The two antennas point toward a distant radio source in a direction indicated by unit vector $\mathbf{s}$. $\mathbf{b}$ is the interferometer baseline, and the wavefront from the source reaches one antenna at a time τ_g later than the other. τ_g is called the *geometrical delay* and is given by

$$\tau_g = \mathbf{b} \cdot \mathbf{s}/c\,, \tag{2–1}$$

where c is the speed of light. The signals from the antennas pass through amplifiers which incorporate filters to select the required frequency band of width $\Delta\nu$ centered on frequency ν. The component in which the signals are combined is the *correlator*, which is a voltage multiplier followed by a time averaging (integrating) circuit. If the input waveforms to the correlator are $V_1(t)$ and $V_2(t)$, the output is proportional to

$$\langle V_1(t) V_2(t) \rangle\,, \tag{2–2}$$

where the angular brackets denote a time average. We can represent the received signals by quasi-monochromatic Fourier components of frequency ν, which have the form $V_1(t) = v_1 \cos 2\pi\nu(t - \tau_g)$ and $V_2(t) = v_2 \cos 2\pi\nu t$. The output of the correlator is then

$$r(\tau_g) = v_1 v_2 \cos 2\pi\nu\tau_g\,. \tag{2–3}$$

τ_g varies slowly with time as the Earth rotates, and the resulting oscillations of the cosine term in Equation 2–3 represent the motion of the source through the interferometer fringe pattern. We may assume that these oscillations are sufficiently slow that the fringes are not significantly attenuated by the averaging (an expression for the fringe frequency is given in Section 8). In contrast, the component at frequency 2ν generated in the multiplication is effectively filtered out. Note that the term $v_1 v_2$, which represents the fringe amplitude, is proportional to the received power.

We now express the interferometer output in terms of the radio brightness integrated over the sky. Let $I(\mathbf{s})$ represent the radio brightness in the direction of unit vector $\mathbf{s}$ at frequency ν. The brightness is also sometimes referred to as intensity and is measured in $\mathrm{W\,m^{-2}\,Hz^{-1}\,sr^{-1}}$. Note that each antenna responds to a component of the input radiation field determined by the antenna polarization. The way in which the antenna polarization is varied to explore the total radiation field is considered in Lectures 3 and 6. The signal power received in bandwidth $\Delta\nu$ from the source element $d\Omega$ is $A(\mathbf{s})I(\mathbf{s})\Delta\nu\, d\Omega$, where $A(\mathbf{s})$ is the effective collecting area in direction $\mathbf{s}$, which we assume to be the same for each of the antennas. The resulting output from the correlator is proportional to the received power and to the cosine fringe term. Thus, omitting constant gain factors, we can represent the correlator output for the signal from solid angle $d\Omega$ by

$$dr = A(\mathbf{s})I(\mathbf{s})\Delta\nu\, d\Omega \cos 2\pi\nu\tau_g\,. \tag{2–4}$$

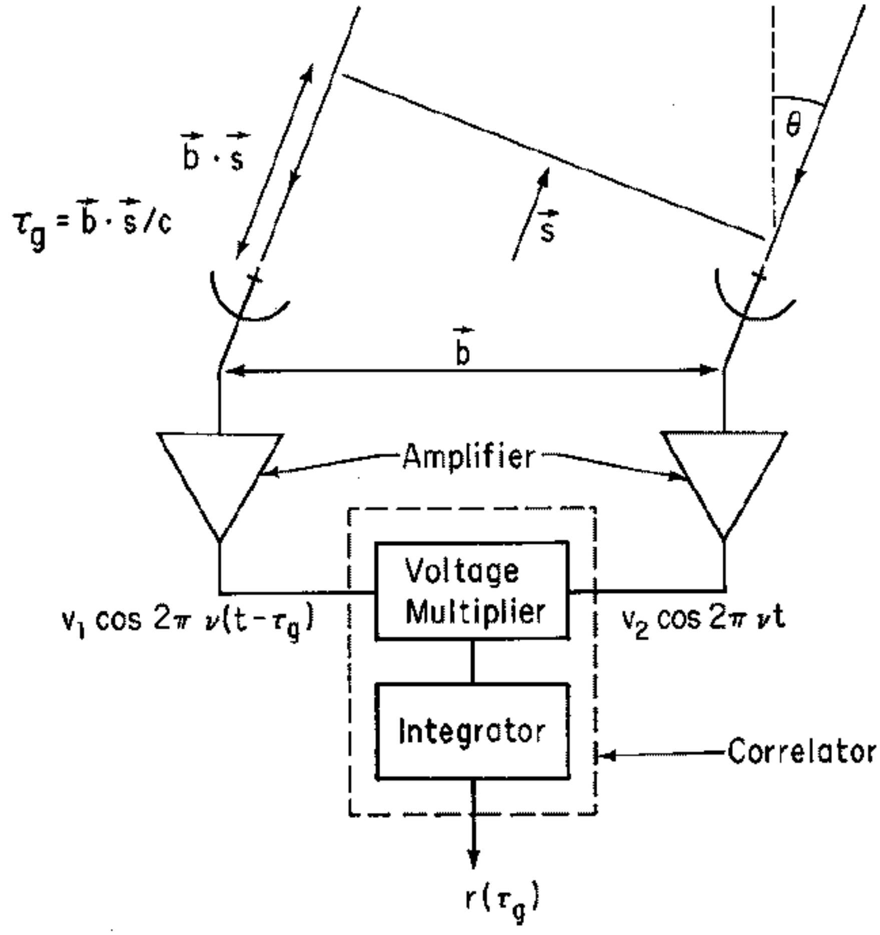


Figure 2–1. Simplified schematic diagram of a two-element interferometer.

In terms of the baseline and source position vectors we can write

$$r = \Delta\nu \int_S A(\mathbf{s})I(\mathbf{s}) \cos \frac{2\pi\nu\, \mathbf{b}\cdot\mathbf{s}}{c}\, d\Omega\,. \tag{2–5}$$

The integral in Equation 2–5 is taken over the entire surface S of the celestial sphere, subtending 4π steradians, but in practice the integrand usually falls to very low values outside a small angular field as a result of the antenna beamwidth, the finite dimensions of the radio source, and other effects which restrict the field of view (see Sections 10 and 11, and Lecture 13). We assume that the bandwidth $\Delta\nu$ is sufficiently small that variation of A and I with ν can be ignored. Two further assumptions have been made in deriving Equation 2–5. First, the source must be in the far field of the interferometer so that the incoming wavefronts can be considered to be plane. With the longest spacings and shortest wavelengths commonly in use, this condition may not be met by some objects within the solar system. Second, the assumption that the responses from different points in the source can be added independently is implicit in the

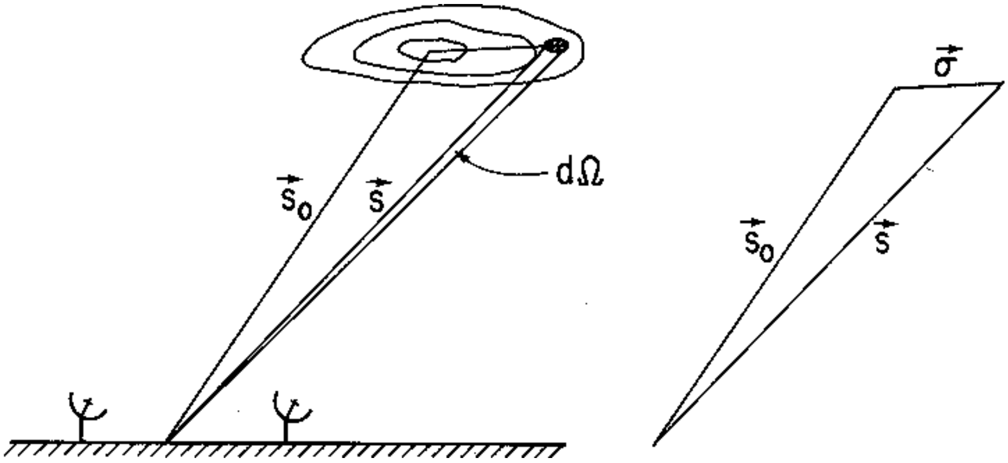


Figure 2–2. Position vectors used in deriving the interferometer response to a source. The source is represented by the contours of radio brightness $I(\mathbf{s})$ on the sky.

integration over angle in Equation 2–5. This requires that the source be spatially incoherent—i.e., that signal components emanating from different points on the source be uncorrelated.

When taking observations to make an interferometric image of a radio source, it is usual to specify a position on which the synthesized field of view is to be centered. This position is commonly referred to as the *phase tracking center* or *phase reference position.* We can represent this position by the vector $\mathbf{s}_0$, as shown in Figure 2–2, and write $\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\sigma}$. From Equation 2–5 we then obtain

$$\begin{aligned} r &= \Delta\nu \cos\left(\frac{2\pi\nu\,\mathbf{b}\cdot\mathbf{s}_0}{c}\right)\int_S A(\boldsymbol{\sigma})I(\boldsymbol{\sigma})\cos\frac{2\pi\nu\,\mathbf{b}\cdot\boldsymbol{\sigma}}{c}\,d\Omega \\ &- \Delta\nu \sin\left(\frac{2\pi\nu\,\mathbf{b}\cdot\mathbf{s}_0}{c}\right)\int_S A(\boldsymbol{\sigma})I(\boldsymbol{\sigma})\sin\frac{2\pi\nu\,\mathbf{b}\cdot\boldsymbol{\sigma}}{c}\,d\Omega\,. \end{aligned} \tag{2–6}$$

We now introduce the *visibility*, which is a measure of the coherence discussed in Lecture 1. The term visibility was first used in interferometry by Michelson (1890) to express the relative amplitude of the optical fringes that he observed. As used in radio astronomy, visibility is a complex quantity, the magnitude of which has the dimensions of spectral power flux density (W m^{-2} Hz^{-1}). It can be regarded as an unnormalized measure of the coherence of the electric field, modified to some extent by the characteristics of the interferometer.

The complex visibility of the source is defined as

$$V \equiv |V|e^{i\phi_V} = \int_S \mathcal{A}(\boldsymbol{\sigma})I(\boldsymbol{\sigma})e^{-2\pi i\nu\mathbf{b}\cdot\,\boldsymbol{\sigma}/c}\,d\Omega \tag{2–7}$$

where $\mathcal{A}(\boldsymbol{\sigma}) \equiv A(\boldsymbol{\sigma})/A_0$ is the normalized antenna reception pattern, A_0 being the response at the beam center. We are considering the case in which the antennas track the source, and the system therefore responds to the modified

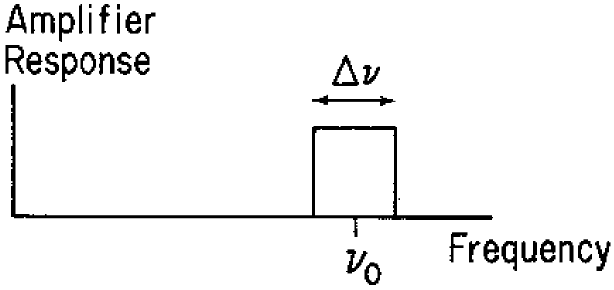


Figure 2–3. Idealized rectangular response of the receiving system.

brightness distribution $\mathcal{A}(\boldsymbol{\sigma})I(\boldsymbol{\sigma})$. By separating the real and imaginary parts of V in Equation 2–7 we obtain

$$A_0|V|\cos\phi_V = \int_S A(\boldsymbol{\sigma})I(\boldsymbol{\sigma})\cos\frac{2\pi\nu\,\mathbf{b}\cdot\boldsymbol{\sigma}}{c}\,d\Omega \qquad (2\text{–}8)$$

and

$$A_0|V|\sin\phi_V = -\int_S A(\boldsymbol{\sigma})I(\boldsymbol{\sigma})\sin\frac{2\pi\nu\,\mathbf{b}\cdot\boldsymbol{\sigma}}{c}\,d\Omega\,. \qquad (2\text{–}9)$$

Substitution of Equations 2–8 and 2–9 into Equation 2–6 gives

$$r = A_0\Delta\nu|V|\cos\left(\frac{2\pi\nu\,\mathbf{b}\cdot\mathbf{s}_0}{c} - \phi_V\right)\,. \qquad (2\text{–}10)$$

In the interpretation of interferometer measurements the usual procedure is to measure the amplitude and phase of the fringe pattern as represented by the cosine term in Equation 2–10, and then derive the amplitude and phase of V by appropriate calibration. The brightness distribution of the source is obtained from the visibility data by inversion of the transformation in Equation 2–7. Thus V must be measured over a sufficiently wide range of $\nu\mathbf{b}\cdot\boldsymbol{\sigma}/c$, which is the component of the baseline normal to the direction of the source and measured in wavelengths. This component can be envisaged as the baseline viewed from the direction of the source.

3. Effect of Bandwidth in a Two-Element Interferometer

Since the frequency of the cosine fringe term in Equation 2–10 is proportional to the observing frequency ν, observing with a finite bandwidth $\Delta\nu$ results, in effect, in the combination of fringe patterns with a corresponding range of fringe frequencies. For the response with an infinitesimal bandwidth $d\nu$ we can write, from Equations 2–1 and 2–10,

$$dr = A_0|V|\cos\left(2\pi\nu\tau_g - \phi_V\right)\,d\nu\,. \qquad (2\text{–}11)$$

Then for a rectangular frequency passband, as shown in Figure 2–3, the interferometer response is

$$r \;=\; A_0|V|\int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2}\cos\left(2\pi\nu\tau_g - \phi_V\right)\,d\nu$$

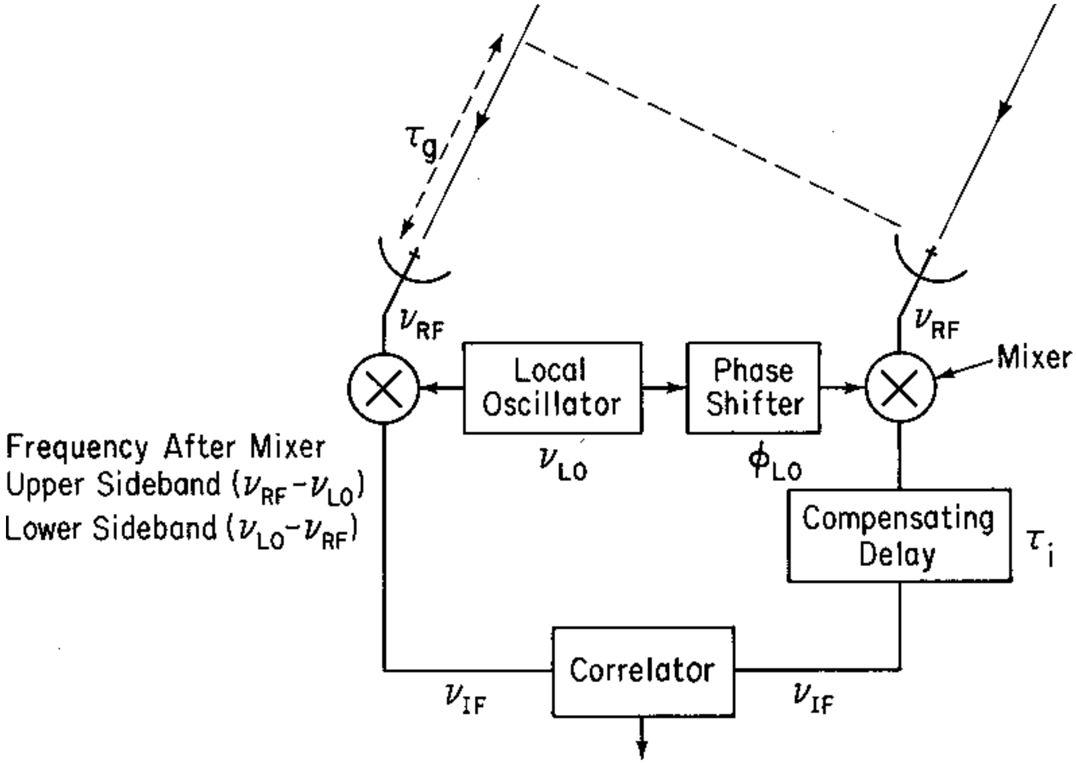


Figure 2–4. Simplified schematic diagram of an interferometer system incorporating frequency conversion and an instrumental time delay to compensate for τ_g. For simplicity, the amplifiers and filters are omitted.

$$= \quad A_0|V|\Delta\nu\frac{\sin\pi\Delta\nu\tau_g}{\pi\Delta\nu\tau_g}\cos\left(2\pi\nu_0\tau_g - \phi_V\right) \ , \qquad (2\text{–}12)$$

where ν_0 is the center frequency of the observing passband. Thus in the system that we are considering the fringes are modulated by a sinc-function envelope, sometimes referred to as the *bandwidth pattern.* The full fringe amplitude is observed only when the source is in a direction normal to the baseline so that $\tau_g = 0$. The range of τ_g for which the fringe amplitude is within, say, 1% of the maximum value can be obtained by writing

$$\frac{\sin\pi\Delta\nu\tau_g}{\pi\Delta\nu\tau_g} \approx 1 - \frac{\left(\pi\Delta\nu\tau_g\right)^2}{6} > 0.99 \qquad (2\text{–}13)$$

which yields $|\Delta\nu\tau_g| < 0.078$, where the approximation in Equation 2–13 is valid for $|\pi\Delta\nu\tau_g| \ll 1$. The angular range of τ_g within this limit depends upon the length and orientation of the baseline: for example, with $\Delta\nu = 50$ MHz and $|\mathbf{b}| = 1$ km, the response falls by 1% when the angle θ in Fig. 2–1 is 2 arcmin. In order to observe a source over a wide range of hour-angle, it is necessary to include within the system a computer-controlled delay to compensate for τ_g.

4. Delay Tracking and Frequency Conversion

A block diagram of an interferometer system that includes an instrumental compensating delay is shown in Figure 2–4. Frequency conversion of the incoming signals at radio frequency ν_{RF} with a local oscillator at frequency ν_{LO} is also included. Practical receiving systems incorporate frequency conversion because it is technically more convenient to perform such functions as amplification, filtering, delaying, and cross-correlating of the signals at an intermediate frequency that is lower than ν_{RF} and remains fixed when the observing frequency is changed. The signals at the frequencies ν_{RF} and ν_{LO} are combined in a mixer which contains a non-linear element (usually a diode) in which combinations of the two frequencies are formed. The intermediate frequency ν_{IF} is related to the mixer input frequencies by

$$\nu_{\mathrm{RF}} = \nu_{\mathrm{LO}} \pm \nu_{\mathrm{IF}}\,. \tag{2–14}$$

Note that ν_{LO} is a single-valued frequency, but ν_{RF} and ν_{IF} refer to bands of width $\Delta\nu$. Thus the mixer responds to inputs in two frequency bands, as shown in Figure 2–5: these are referred to as the upper and lower sidebands and correspond to the $+$ and $-$ signs in Equation 2–14 respectively. For observations at frequencies up to a few tens of gigahertz the signal from each antenna is usually first applied to a low-noise amplifier to obtain high sensitivity, and then passed through a filter that transmits only one of the two sidebands to the mixer. The response of such a single-sideband system can be obtained by considering the phase changes ϕ_1 and ϕ_2 imposed upon the signals received by antennas 1 and 2 before reaching the correlator inputs. For the upper sideband case we have

$$\begin{aligned}\phi_1 &= 2\pi\nu_{\mathrm{RF}}\tau_g = 2\pi(\nu_{\mathrm{LO}} + \nu_{\mathrm{IF}})\tau_g\,,\\ \phi_2 &= 2\pi\nu_{\mathrm{IF}}\tau_i + \phi_{\mathrm{LO}}\,,\end{aligned} \tag{2–15}$$

where ϕ_{LO} is the difference in the phase of the local oscillator signal at the two mixers, and τ_i is the instrumental delay that compensates for τ_g. The upper-sideband response of the interferometer is obtained by replacing the argument of the cosine function in Equation 2–11 by $\phi_1 - \phi_2 - -\phi_V$, $d\nu$ by $d\nu_{\mathrm{IF}}$, and integrating with respect to ν_{IF} from $\nu_{\mathrm{IF}_0} - \Delta\nu/2$ to $\nu_{\mathrm{IF}_0} + \Delta\nu/2$. Thus:

$$r_u = A_0\Delta\nu|V|\frac{\sin\pi\Delta\nu\Delta\tau}{\pi\Delta\nu\Delta\tau}\cos[2\pi(\nu_{\mathrm{LO}}\tau_g + \nu_{\mathrm{IF}_0}\Delta\tau) - \phi_V - \phi_{\mathrm{LO}}]\,. \tag{2–16}$$

Here $\Delta\tau = \tau_g - \tau_i$ is the tracking error of the compensating delay τ_i. Note that the output fringe oscillations, which result from the time variation of τ_g, in this case depend upon the local oscillator frequency ν_{LO} rather than the observing frequency at the antenna as in Equation 2–10. For the case in which the lower sideband is the one that is accepted by the receiving system we have:

$$\begin{aligned}\phi_1 &= -2\pi(\nu_{\mathrm{LO}} - \nu_{\mathrm{IF}})\tau_g\,,\\ \phi_2 &= 2\pi\nu_{\mathrm{IF}}\tau_i - \phi_{\mathrm{LO}}\,,\end{aligned} \tag{2–17}$$

whence

$$r_l = A_0\Delta\nu|V|\frac{\sin\pi\Delta\nu\Delta\tau}{\pi\Delta\nu\Delta\tau}\cos[2\pi(\nu_{\mathrm{LO}}\tau_g - \nu_{\mathrm{IF}_0}\Delta\tau) - \phi_V - \phi_{\mathrm{LO}}]\,. \tag{2–18}$$

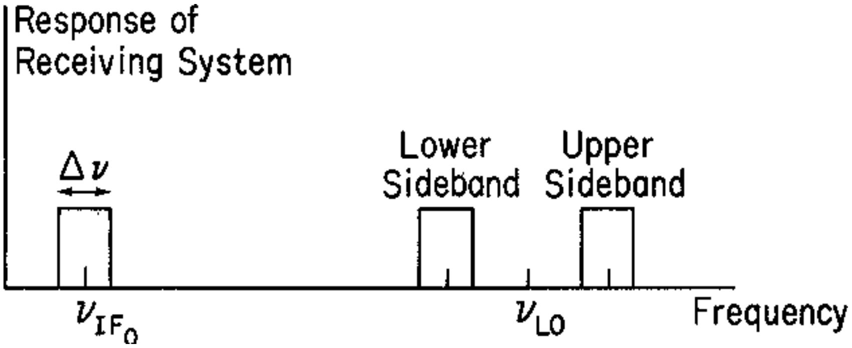


Figure 2–5. Relationship of RF (upper and lower sideband), IF, and LO frequencies.

Here the differences in the signs of the various terms compared with those in Equation 2–15 occur because in lower sideband conversion a change in phase of the RF signal causes a phase change of opposite sign in the IF signal. The phase of the local oscillator also enters with a different sign in Equation 2–15 and 2–17.

At frequencies approaching 100 GHz and higher, it is difficult to make low-noise amplifiers to place ahead of the mixers. Often the antenna is connected directly to the mixer input, without any filter to reject one sideband since such a filter can introduce noise unless cryogenically cooled. The result is a double-sideband system, and the response is obtained from the sum of Equation 2–16 and 2–18:

$$\begin{aligned} r_d &= r_u + r_l \qquad\qquad (2\text{–}19) \\ &= 2\Delta\nu A_0|V|\frac{\sin(\pi\Delta\nu\Delta\tau)}{\pi\Delta\nu\Delta\tau}\cos(2\pi\nu_{\mathrm{LO}}\tau_g - \phi_V - \phi_{\mathrm{LO}})\cos(2\pi\nu_{\mathrm{IF}_0}\Delta\tau)\,. \end{aligned}$$

Note that the delay-tracking error $\Delta\tau$ does not affect the phase of the cosine fringe term as it does in Equation 2–16 and 2–18, but here it appears in a separate cosine term that modulates the amplitude of the fringes. As a result, the double-sideband system requires more critical adjustment of the instrumental delay to maintain the visibility amplitude than does the single-sideband system. Other disadvantages of the double-sideband system include greater vulnerability to interference, and complication of spectral line observations since the spectra of the two sidebands are superimposed. Separation of the sideband responses after correlation of the signals by a technique involving periodic insertion of $\pi/2$ phase shifts in the local oscillator is used in some instruments: for a discussion see Thompson, Moran & Swenson (1986).

5. Fringe Rotation and Complex Correlators

The output from the correlator represented by Equation 2–16, 18 or 19 is fed to a computer which performs some form of optimal analysis to determine the amplitude and phase of the fringe oscillations. The fringe visibility V can then be obtained by calibration of the instrumental parameters. This calibration usually involves observation of one or more sources with known positions, flux

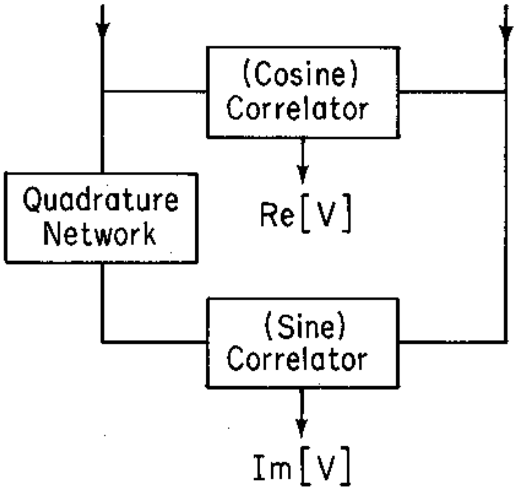


Figure 2–6. Complex correlator system. The quadrature network introduces a $\pi/2$ phase shift: a signal of the form $\cos 2\pi\nu t$ at its input becomes $\cos(2\pi\nu t - \pi/2)$ at the output.

densities, and angular dimensions. For an array such as the VLA, the frequencies of the fringe oscillations can exceed 150 Hz for the longest antenna spacings, and in VLBI the fringe frequency can exceed 100 kHz. To preserve the fringe information it is necessary to sample the correlator output at least twice per fringe period. Thus the data rate to the computer can be very much higher than that necessary to follow the changes in the visibility V, for which values at intervals of order one second are likely to be adequate. However, by inserting progressively varying phase shifts in the local oscillator signals it is possible to slow down the fringe oscillations, and reduce the computation required. Thus in Equation 2–16, 2–18 and 2–19, if we vary $\phi_{\rm LO}$ so that $(2\pi\nu_{\rm LO}\tau_g - \phi_{\rm LO})$ remains constant, the correlator output will vary only as a result of changes in V and slow drifts in the instrumental parameters. This procedure, in which $\phi_{\rm LO}$ is usually controlled by the same computer that regulates the delay tracking, is variously referred to as *fringe rotation* or *fringe stopping.* Note that the effect of the compensating delay τ_i tracking τ_g is to cause the envelope of the fringe pattern to follow the source across the sky, and to change the frequency of the fringes by a factor $\nu_{\rm LO}/\nu_0$ for a single-sideband receiving system. If τ_i were inserted at the received signal frequency, the fringe frequency would be reduced to zero without adjustment of the local oscillator phase.

After fringe stopping, the output of the correlator in Figure 2–4 is a slowly varying voltage (a constant voltage for the case of a point source at the phase tracking center). To measure the complex fringe amplitude in this case, a scheme using two correlators, as shown in Figure 2–5 can be used. For each antenna pair a second correlator with a $\pi/2$ phase shift in one input is added. The response of the second correlator can be obtained by replacing ϕ_1 in Equations 2–15 and 2–17 by $\phi_1 - \pi/2$. Then in Equations 2–16, 2–18 and 2–19 the cosine term containing τ_g becomes a sine, with no change in the argument. The two

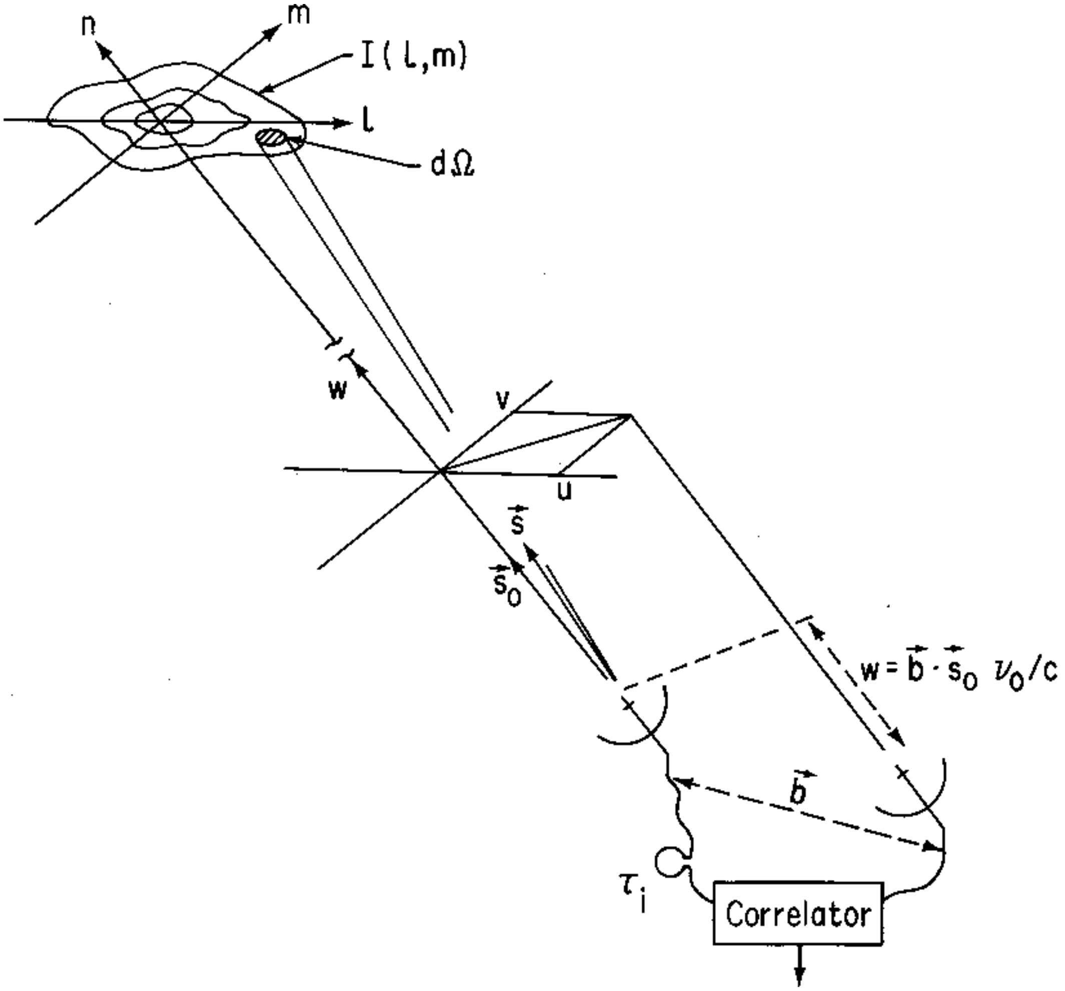


Figure 2–7. The (u, v, w) and (l, m, n) right-handed coordinate systems used to express the interferometer baselines and the source brightness distribution, respectively.

outputs in Figure 2–6 can thus be regarded as measuring the real and imaginary parts of the complex fringe amplitude, or complex visibility. Such a scheme is usually referred to as a *complex correlator.* In addition to allowing the visibility to be measured with zero fringe frequency, the complex correlator provides an improvement of $\sqrt{2}$ in signal-to-noise ratio over a single correlator, since the noise fluctuations at the two outputs are uncorrelated. See Lecture 9 for an analysis of signal-to-noise ratios.

6. Phase Switching

Phase switching is a technique that is included in many interferometer systems to eliminate errors in the form of constant or slowly varying offsets in the correlator outputs. Such errors can result from misadjustment of the correlator circuitry, cross coupling between the signals at the correlator inputs, and various other effects. They can be very effectively reduced by periodically reversing the phase of one of the signals at an early point in the receiving system, and synchronously reversing the sign of the multiplier output in the correlator, before the data are averaged. For the wanted component of the signal, the two reversals cancel one another, but unwanted components in the multiplier output which do not reverse sign with reversal of the phase of a received signal are averaged towards zero. In practice, the frequency of the switching is of the order of 10 or 100 Hz. This technique, known as phase switching, was first introduced by Ryle (1952) as a means of implementing the multiplicative action of a correlator using a power-linear diode detector. For a description of a more recent application of phase switching see Granlund, Thompson & Clark (1978).

7. Coordinate Systems for Imaging

The practical application of Equation 2–7 requires the introduction of a coordinate system, and the one that is usually chosen was introduced in Lecture 1 and is shown in Figure 2–7. The baseline vector has components (u, v, w) where w points in the direction of interest, i.e., towards a position $\mathbf{s}_0$ that becomes the center of the synthesized image. Note that u, v, and w are measured in wavelengths at the center frequency of the RF signal band, and in directions towards the East, the North, and the phase tracking center, respectively. Positions on the sky are defined in l and m, which are direction cosines measured with respect to the u and v axes. A synthesized image in the (l, m) plane represents a projection of the celestial sphere onto a tangent plane at the (l, m) origin. Distances in l and m are proportional to the sines of the angles measured from the origin, which is a convenient practical system. In these coordinates the parameters used in the derivation of the interferometer response in terms of visibility (Eqs. 6 and 7) become

$$\begin{aligned} \frac{\nu\,\mathbf{b}\cdot\mathbf{s}}{c} &= ul + vm + wn, \\ \frac{\nu\,\mathbf{b}\cdot\mathbf{s}_0}{c} &= w, \\ \text{and} \qquad d\Omega = \frac{dl\,dm}{n} &= \frac{dl\,dm}{\sqrt{1-l^2-m^2}}. \end{aligned} \tag{2–20}$$

Thus in the coordinates of Figure 2–7, Equation 2–7 becomes

$$V(u,v,w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}(l,m) I(l,m) e^{-2\pi i\left[ul+vm+w\left(\sqrt{1-l^2-m^2}-1\right)\right]} \frac{dl\,dm}{\sqrt{1-l^2-m^2}}, \tag{2–21}$$

where the integrand is taken to be zero for $l^2 + m^2 \geq 1$. Note that we express the complex visibility as a function of (u, v, w), since these are the coordinates that represent the spacings of the antennas with respect to the phase tracking center of the source, $\mathbf{s}_0$. The visibility is also a function of the modified brightness distribution $\mathcal{A}I$.

To simplify the inversion of Equation 2–21, by means of which $I(l, m)$ is obtained from the visibility, it is desirable to reduce this equation to the form

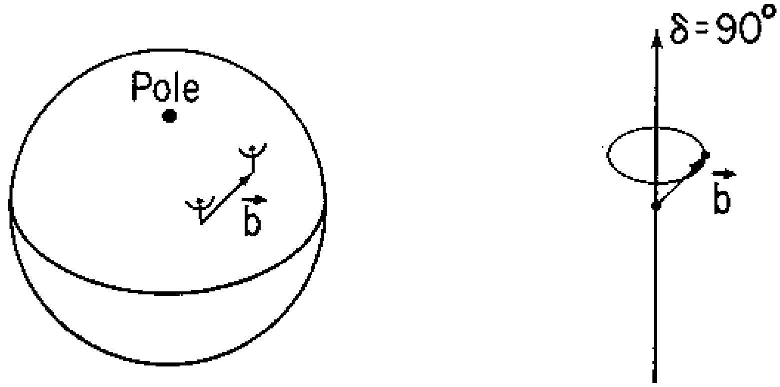


Figure 2–8. As the Earth rotates, the baseline vector **b**, which represents the spacing of the two antennas, traces out a circular locus in a plane normal to the direction of declination (δ) equal to 90°. If the antennas are in an East–West line on the Earth, then the vector **b** is normal to the rotation axis.

of a two-dimensional Fourier transform. This can be done under two sets of conditions. The first is when the baselines are coplanar, which can be understood by considering the way in which the Earth's rotation carries the antennas through space. It should be evident from Figure 2–8 that the rotation causes the tip of the baseline vector to trace out a circle concentric with the Earth's rotation axis. The rising and setting of a point on the sky usually limit the range over which V can be measured to an arc of the circle. In general, for a two-dimensional array of antennas on the surface of the Earth, the circular loci resulting from the different baselines have different diameters and lie in different planes. However, for the particular case of an array of antennas in an East–West line on the Earth's surface, the components of the baseline vector parallel to the Earth's axis are zero, and the baseline-vectors are coplanar. Then, if we choose the w-axis to lie in the direction of the celestial pole, so that $w = 0$, Equation 2–21 becomes

$$V(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}(l,m)I(l,m)e^{-2\pi i(ul+vm)}\frac{dl\,dm}{\sqrt{1-l^2-m^2}}\,. \qquad (2\text{–}22)$$

This equation is a two-dimensional Fourier transform, the inverse of which is

$$\frac{\mathcal{A}(l,m)I(l,m)}{\sqrt{1-l^2-m^2}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \qquad (2\text{–}23)$$

Equation 2–23 can be applied to all parts of the hemisphere shown in Figure 2–9. Usually we want to image a small area of the sky defined by the antenna beams. If this is centered on right ascension α_0 and declination δ_0, we can choose the direction of the u-axis as in Figure 2–9 so that l is small within the region of interest and is approximately equal to angular distance on the sky. However, m remains the sine of the angular distance measured from the pole, i.e., $m = \cos\delta$, and the scale of the image is compressed in the m-direction by a factor $\sin\delta$. Thus one would prefer to center the image on (α_0, δ_0). This can be done by rotating the axes about the u-direction until w points towards (α_0, δ_0)

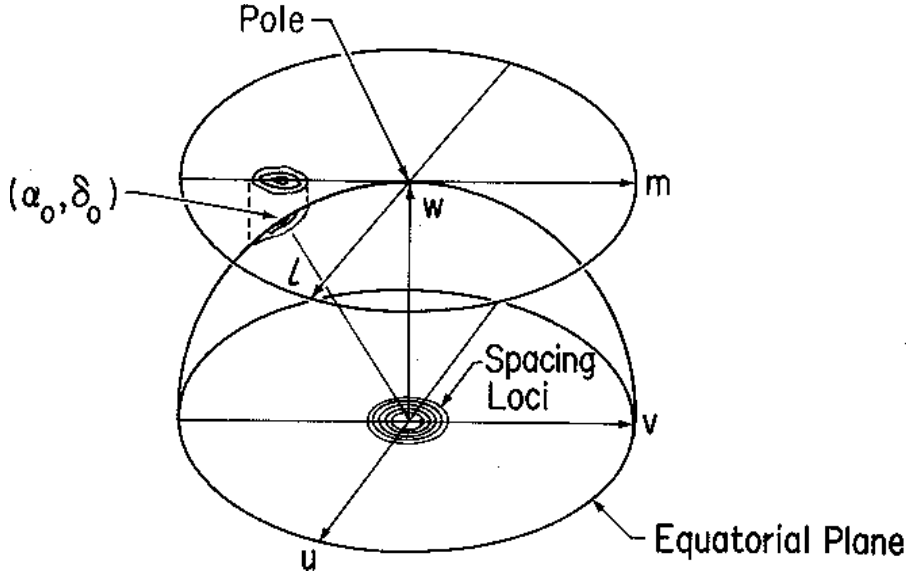


Figure 2–9. Celestial hemisphere showing the projection of a source at (α_0, δ_0) onto the tangent plane at the pole. The spacing-vector loci are for an array with East–West baselines, and lie in a plane parallel to the Earth's equator. The direction of the w-axis is here chosen to be that of the pole ($\delta = 90°$).

and substituting in Equation 2–21 $w = -v\cos\delta_0$, which follows from the location of the baselines in the equatorial plane. Then m can be redefined to provide a two-dimensional Fourier transform relationship. This procedure is required only for large field images, and in most cases the small field approximation, which will be described next, suffices.

It is clear from Figure 2–9 that for an East–West array the projected spacings of the antenna pairs become seriously foreshortened in the v-direction for the observations at low declinations. In that part of the sky it is necessary to use baselines with a significant component parallel to the Earth's axis, i.e., non-East–West baselines. Thus for a two-dimensional array of antennas the baseline vectors do not remain coplanar in (u, v, w) space. A system of three coordinates is required to accommodate the spacing vectors, and we must consider the second set of conditions under which Equation 2–21 reduces to a two-dimensional Fourier transform. These depend upon $|l|$ and $|m|$ being small enough that we can write

$$\left(\sqrt{1-l^2-m^2}-1\right)w \approx -\frac{1}{2}(l^2+m^2)w \approx 0\,. \tag{2–24}$$

Then Equation 2–21 becomes

$$V(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}(l,m)I(l,m)e^{-2\pi i(ul+vm)}\,dl\,dm\,. \tag{2–25}$$

For $|l|$ and $|m|$ small, i.e., small field imaging, the dependence of the visibility upon w is very small and can be omitted. From Equation 2–25 we can write

$$\mathcal{A}(l,m)I(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{2–26}$$

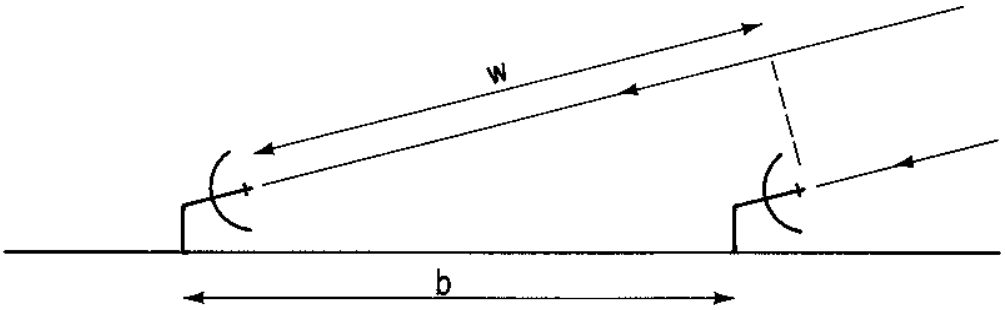


Figure 2–10. Comparison of the w-component and the antenna spacing when the direction of the source is close to that of the baseline. This condition can occur when the source is rising or setting.

For arrays in which the baselines do not remain coplanar as the Earth rotates, the approximation in Equation 2–24 results in a phase error of $\pi(l^2+m^2)w$ for radiation from the point (l, m). Note that the condition for the approximation in Equation 2–24 to be valid is $|\pi(l^2+m^2)w| \ll 1$, not just $l^2+m^2 \ll 1$. Unless special procedures are used, this condition places a limit on the size of the source that can be imaged without distortion. The limit can be roughly estimated as follows: For any pair of antennas the maximum value of w occurs when the source under observation is at a low angle of elevation and an azimuth close to that of the baseline, as shown in Figure 2–10. Under such circumstances w is approximately equal to b/λ, the baseline length measured in wavelengths. Thus for an array of antennas for which the half-power width of the synthesized beam is $\theta_{\rm HPBW}$, we can write

$$\frac{1}{\theta_{\rm HPBW}} \approx \frac{b_{\rm max}}{\lambda} \approx w_{\rm max}\,, \tag{2–27}$$

where $b_{\rm max}$ is the longest baseline. If $\theta_{\rm F}$ is the width of the synthesized field, the maximum phase error is about

$$\frac{\pi\theta_{\rm F}^2}{4\theta_{\rm HPBW}}\,. \tag{2–28}$$

Since this is the *maximum* phase error, we can possibly allow it to be as high as 0.1 radian without introducing serious errors in the image, from which we obtain

$$\theta_{\rm F} < \frac{1}{3}\sqrt{\theta_{\rm HPBW}}\,, \tag{2–29}$$

where the two angles are measured in radians. Then, for example, if $\theta_{\rm HPBW} = 2''$, $\theta_{\rm F} < 5'$. For fields of greater width than allowed by Equation 2–29 there are ways of avoiding or ameliorating the distortion introduced by the phase errors, at the expense of more complicated algorithms and increased computing—see Lecture 19.

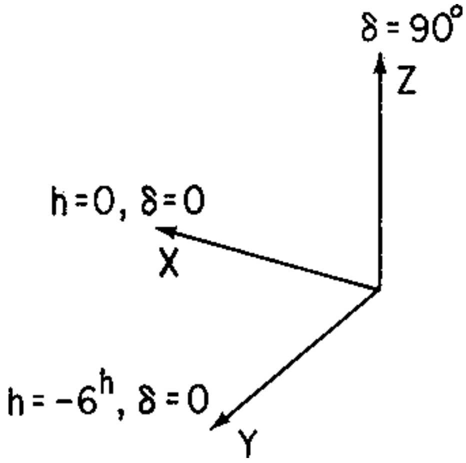


Figure 2–11. Coordinate system for specification of baseline parameters. X is the direction of the meridian at the celestial equator, Y is toward the East, and Z toward the North celestial pole.

8. Antenna Spacings and (u, v, w) Components

With multiple-element antenna arrays, it is convenient to specify the antenna positions relative to some reference point measured in a Cartesian coordinate system. For example, a system with axes pointing towards hour-angle h and declination δ equal to $(h = 0, \delta = 0)$ for X, $(h = -6^{\rm h}, \delta = 0)$ for Y, and $(\delta = 90°)$ for Z may be used as in Figure 2–11. Then if L_X, L_Y, and L_Z are the corresponding coordinate differences for two antennas, the baseline components (u, v, w) are given by

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} \sin H_0 & \cos H_0 & 0 \\ -\sin\delta_0 \cos H_0 & \sin\delta_0 \sin H_0 & \cos\delta_0 \\ \cos\delta_0 \cos H_0 & -\cos\delta_0 \sin H_0 & \sin\delta_0 \end{pmatrix} \begin{pmatrix} L_X \\ L_Y \\ L_Z \end{pmatrix}, \qquad (2\text{–}30)$$

where H_0 and δ_0 are the hour-angle and declination of the phase reference position, and λ is the wavelength corresponding to the center frequency of the receiving system. The elements in the transformation matrix in Equation 2–30 are the direction cosines of the (u, v, w) axes relative to (X, Y, Z) axes: for further details see, e.g., Thompson, Moran & Swenson (1986). By eliminating H_0 from the expressions for u and v we obtain the equation of an ellipse in the (u, v) plane:

$$u^2 + \left(\frac{v - (L_Z/\lambda)\cos\delta_0}{\sin\delta_0} \right)^2 = \frac{L_X^2 + L_Y^2}{\lambda^2} . \qquad (2\text{–}31)$$

Thus as the interferometer observes a point on the celestial sphere, the rotation of the Earth causes the u and v components of the baseline to trace out an elliptical locus. This ellipse is simply the projection onto the (u, v) plane of the circular locus traced out by the tip of the baseline vector, as shown earlier in Figure 2–8. Since $I(l, m)$ is real, $V(-u, -v) = V^*(u, v)$, and at any instant the correlator output provides a measure of the visibility at two points in the (u, v)

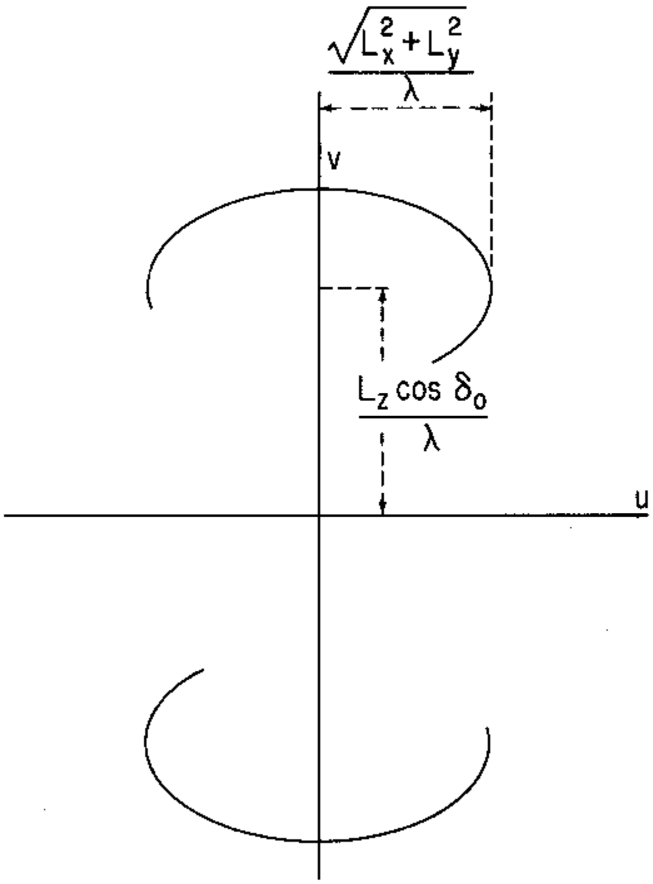

$\frac{\sqrt{L_x^2+L_y^2}}{\lambda}$
v
$\frac{L_z \cos \delta_0}{\lambda}$
u

Figure 2–12. Elliptical loci representing the projection of the baseline vector onto the (u, v) plane as a source is tracked across the sky. The lower curve corresponds to a reversal of the direction of the baseline vector, and represents the points for which the visibility is the complex conjugate of that measured on the upper curve. The axial ratio of the ellipses is equal to $\sin \delta_0$. For an East–West baseline $L_Z = 0$, and a single ellipse is centered on the (u, v) origin.

plane, as in Figure 2–12. For an array of antennas the ensemble of elliptical loci is known as the *transfer function* or *sampling function*, $S(u, v)$, which is a function of the declination of the observation as well as of the antenna spacings. The transfer function indicates the values of u and v at which the visibility function is sampled. Since the visibility function for a point source at the (l, m) origin is a constant in u and v, the Fourier transform of the transfer function indicates the response to a point source, i.e., the synthesized beam. In designing arrays the principal aim is to obtain transfer functions that cover the (u, v) plane as widely and as uniformly as possible. The term transfer function was introduced from an analogy with electrical filter theory. An interferometer responds to structure on the sky with spatial frequency u cycles per radian in the l-direction and v cycles per radian in the m-direction. The transfer function of an array therefore indicates its response as a spatial frequency filter.

9. Astronomical Data from Interferometer Observations

In synthesis imaging an interferometer or array is used to provide values of the complex visibility as a function of u and v, from which a brightness distribution can be derived. For this purpose the visibility measurements should be fairly uniformly distributed over the (u, v) plane, from the origin to some outer boundary that determines the angular resolution. The design of synthesis arrays, which we discuss below, is based largely upon these considerations. If, however, we wish to measure the positions of a series of unresolved sources, the principal consideration is the ability to interpolate the measured visibility phase between one baseline and another, and uniformity of coverage is less important. This consideration also applies to measurements used to monitor universal time, polar motion and geodynamic variation in antenna positions.

In addition to the measurement of complex visibility, two other characteristics of the interferometer output can be used to determine astronomical data. These are principally of importance in VLBI, in which it is not always possible to calibrate the interferometer fringe phase. The first is the bandwidth pattern in Equation 2–12, which can be used to measure τ_g. This is accomplished by finding the value of the instrumental delay τ_i that maximizes the fringe amplitude. A wide receiver bandwidth, or a series of narrow bands at different frequencies, is used to minimize the width of the fringe envelope as a function of τ_i and thereby increase the accuracy. For a source at position (H_0, δ_0), τ_g is equal to w/ν_0 where w is given by Equation 2–30. The second characteristic that can be measured is the fringe frequency. Since the relative phase of the signal at the two antennas changes by 2π when w changes by one (wavelength), the fringe frequency is equal to dw/dt, which can be obtained from Equation 2–30 by differentiation. A useful expression for the fringe frequency ν_F is

$$\nu_F = \frac{dw}{dt} = -\omega_e u \cos\delta \,, \tag{2–32}$$

where $\omega_e = dH_0/dt$ is the angular rotation velocity of the Earth. Equation 2–32 applies when the instrumental delay τ_i is held constant. When τ_i tracks τ_g a factor ν_0/ν_{LO}, ν_0 being the center of the receiving band, should be included for single-sideband receiving systems. In either case, ν_F goes through zero on the v-axis of the (u, v) plane. Note that a single observation of w and dw/dt is sufficient to determine the position of a source if the interferometer baseline is known.

10. Design of Synthesis Arrays

In an array of n_a antennas, a total of $\frac{1}{2}n_a(n_a - 1)$ pair combinations can be formed. The signal from each antenna is then divided in $n_a - 1$ ways and fed to a system of correlators. The rate at which visibility measurements can be made, relative to that for a single interferometer, is approximately proportional to n_a^2. Note that since the signals are amplified before splitting there is no loss in sensitivity, as may occur in instruments for infrared or shorter wavelengths. The primary concern in designing the configuration of antennas is to obtain coverage of the (u, v) plane (i.e., sampling of the visibility function) as uniformly and efficiently as possible over a range determined by the required angular resolution.

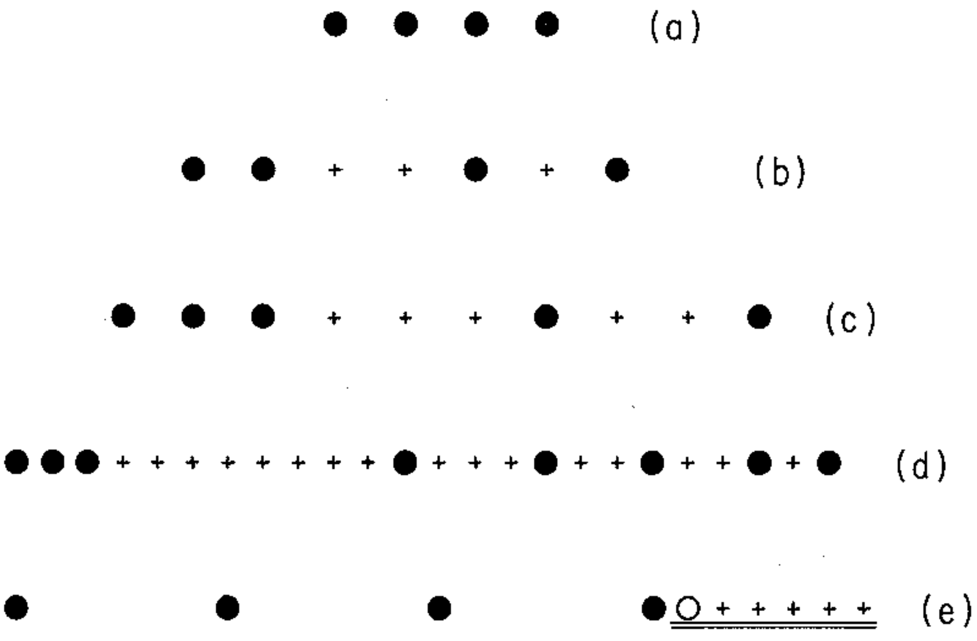


Figure 2–13. Examples of several types of linear arrays of antennas. **(a)** Uniform-spacing array, **(b)** non-redundant array (Arsac 1955), **(c)** minimum-redundancy array (Bracewell 1966), **(d)** minimum-redundancy array (Moffet 1968), and **(e)** array with movable element represented by the open circle.

A commonly used configuration of antennas for synthesis imaging is an East–West linear array. If the various pair combinations of the antennas encompass a series of spacings which increase by a constant increment, the transfer function consists of a series of ellipses centered on the (u, v) origin with a constant increment in the major axes. The axial ratios of the ellipses are equal to $\sin \delta_0$, as in Figure 2–12, which largely determines the axial ratio of the synthesized beam. Thus, for angular distances greater than about 30° from the celestial equator, East–West linear arrays are satisfactory for two-dimensional imaging. Some basic considerations of linear configurations of antennas are illustrated in Figure 2–13. In a simple, uniformly-spaced array as in (a) the longest spacing is $n_a - 1$ times the unit spacing. The shorter spacings occur more than once and are highly redundant. Figure 2–13(b) shows a non-redundant arrangement of four antennas designed by Arsac (1955). For more than four antennas there is always some redundancy, as in the example by Bracewell (1966; see also Bracewell *et al.* 1973) in Figure 2–13(c). Other examples of minimum-redundancy arrays are described by Moffet (1968), and an example with eight antennas for which the longest spacing is 23 times the unit spacing is shown in Figure 2–13(d). Only a few such arrays have been constructed for radio astronomy, and configurations with a number of movable antennas, which offer greater flexibility, are generally preferred (see Lecture 27).

Figure 2–13(d) shows an arrangement of four fixed antennas and one movable one. By repeating an observation for each position of the movable antenna, as indicated by the crosses, it is possible to include all baselines up to the overall length of the array, with intervals equal to the increments in the position of the movable antenna. Although several days are required to complete an observation, a large number of baselines can be covered using a relatively small number of antennas, and highly detailed images obtained. A number of notable instruments have been designed using this principle: these include the One–Mile and Five–Kilometer arrays at Cambridge (Ryle 1962, 1972), the Westerbork Synthesis Radio Telescope (Högbom & Brouw 1974), and the Australia Telescope (Frater 1984). For observations at low declinations, two-dimensional configurations of antennas are generally required to obtain adequate resolution in both right ascension and declination. The design of two-dimensional arrays is more of an empirical matter than that of one-dimensional arrays, since there are no known solutions similar to those based on variability of location of small numbers of antennas or on minimum-redundancy. The main concern is to obtain adequate coverage of the (u, v) plane, whilst using a fairly simple geometrical configuration for reasons of economy. These considerations are illustrated by the design of the VLA (Thompson *et al.* 1980; Napier, Thompson & Ekers 1983). The antenna configuration and examples of the transfer function for the VLA are shown in Figure 2–14 In the configuration in Figure 2–14a the distance from the center of the array of the n^{th} antenna on each arm, counting outwards from the center, is proportional to $n^{1.716}$. With this power-law design, no two spacings on any arm are equal. The array is rotated through 5° from the position of North–South symmetry to avoid exact East–West baselines, which would otherwise occur between antennas on the two Southern arms. At declination 0° the (u, v) components for all East–West baselines become coincident with the u-axis. Thus the power-law spacing and the rotation are features of the VLA design that reduce redundancy in the coverage of the (u, v) plane.

The same considerations of uniformity of sampling in the (u, v) plane also apply to arrays for imaging by VLBI. The main practical difference is that since the antennas are not directly interconnected, except by telephone lines for monitor and control purposes, there is no advantage to any particular geometric pattern. Thus, after the (u, v) coverage, the main concern is the choice of sites for freedom from interference, low water vapor in the atmosphere, convenience for service, etc. The locations for antennas in the Very Long Baseline Array (VLBA) (Napier *et al.* 1994), and examples of transfer functions, are shown in Figure 2–15. An example of the effect of the addition of the low Earth orbit satellite HALCA to the VLBA is shown in Figure 2–16. The orbital motion significantly increases the resolution along one axis, but at the cost of large holes in the (u, v) plane. For even longer spacings, it would be possible to use two or more antennas in higher orbits, with periods differing by about 10%, to give a wide distribution of spacings (Preston *et al.* 1983). The subject of orbiting VLBI is more fully discussed in Lecture 26.

11. The Effect of Bandwidth in Radio Images

We have seen in Section 2 that the effect of a finite receiving bandwidth $\Delta\nu$ is to modulate the fringes with an envelope function of width inversely proportional to $\Delta\nu$, and that as a result we must insert an instrumental delay τ_i to compensate for the geometrical delay τ_g. This compensation is exact only for radiation from

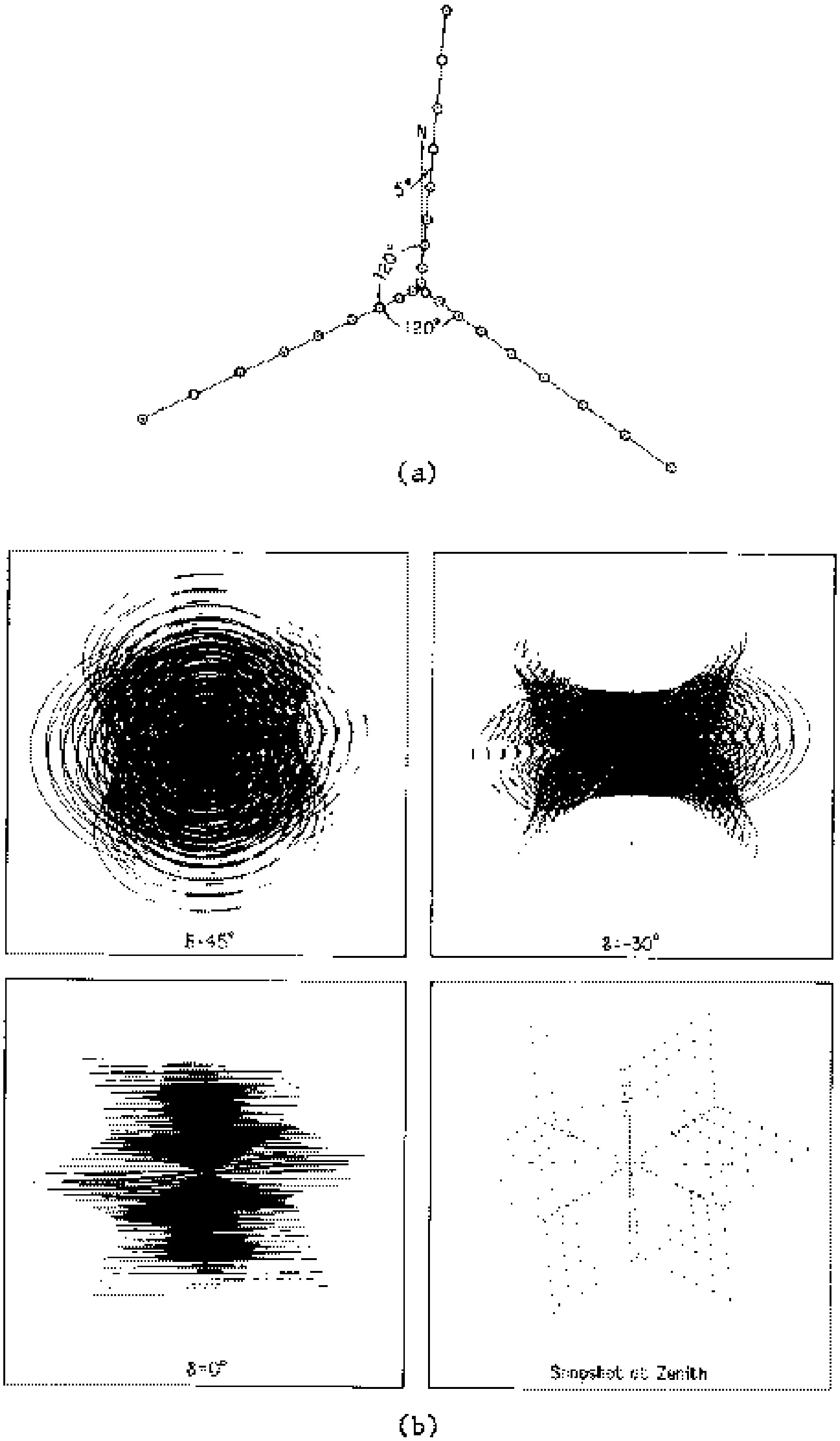


Figure 2–14. **(a)** The configuration of the 27 antennas of the VLA. **(b)** The transfer functions for four declinations with observing durations of ±4^h for $\delta = 0°$ and 45°, ±3^h for $\delta = -30°$, and ±5^m for the snapshot. [From Napier, Thompson & Ekers (© 1983 IEEE).]

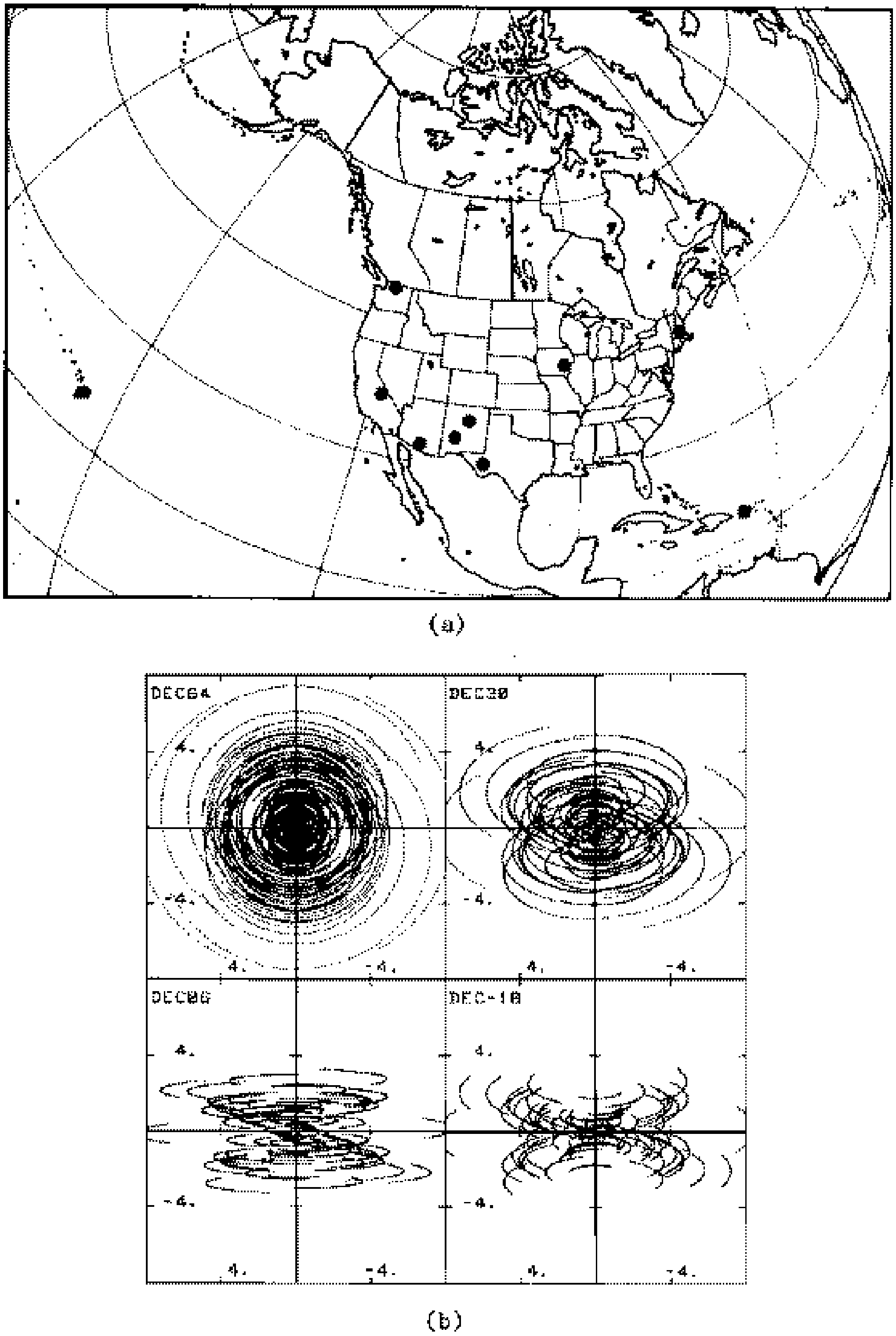


Figure 2–15. **(a)** Locations of the ten antennas of the VLBA, as shown by the closed circles. **(b)** The corresponding transfer functions for four declinations. [From Walker 1984.]

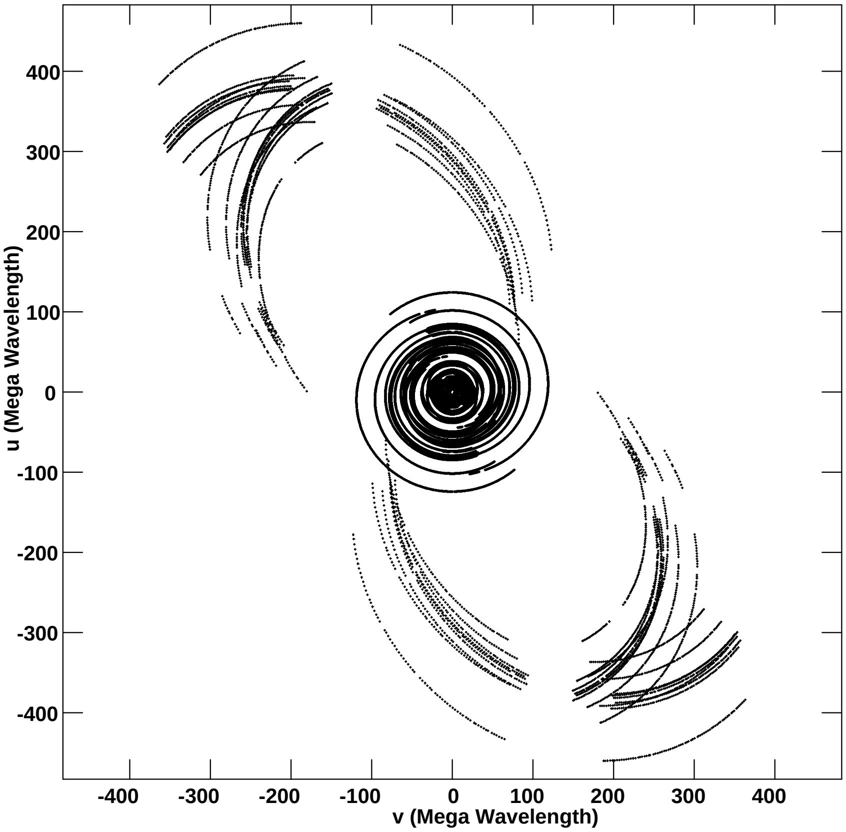


Figure 2–16. The (u,v) coverage provided with the low Earth orbit VLBI satellite HALCA and the ten antennas comprising the VLBA. Three satellite passes are included. The observations were of the northern object 0212+735 at a wavelength of 6cm.

the center of the synthesized field, which is usually chosen as the delay tracking center. Variation of τ_g over the field causes a radial blurring of the image (see, e.g., Thompson & D'Addario 1982), as will now be described.

In observing continuum radiation one is interested in the mean brightness over the bandwidth $\Delta\nu$, and the visibility data are processed as though they were all observed at the center frequency ν_0 indicated in Figure 2–17a. In particular, the spatial frequency coordinates in the (u,v) plane are calculated for the band center. Let these be (u_0,v_0) for frequency ν_0 and (u,v) for another frequency ν within the receiving band. Since u and v represent projected antenna spacings measured in wavelengths, we can write

$$(u_0,v_0) = \left(\frac{\nu_0}{\nu}u, \frac{\nu_0}{\nu}v\right) . \qquad (2\text{–}33)$$

Now consider the visibility that corresponds to a very small band of frequencies centered on ν as in Figure 2–17(a). This band contributes a component of

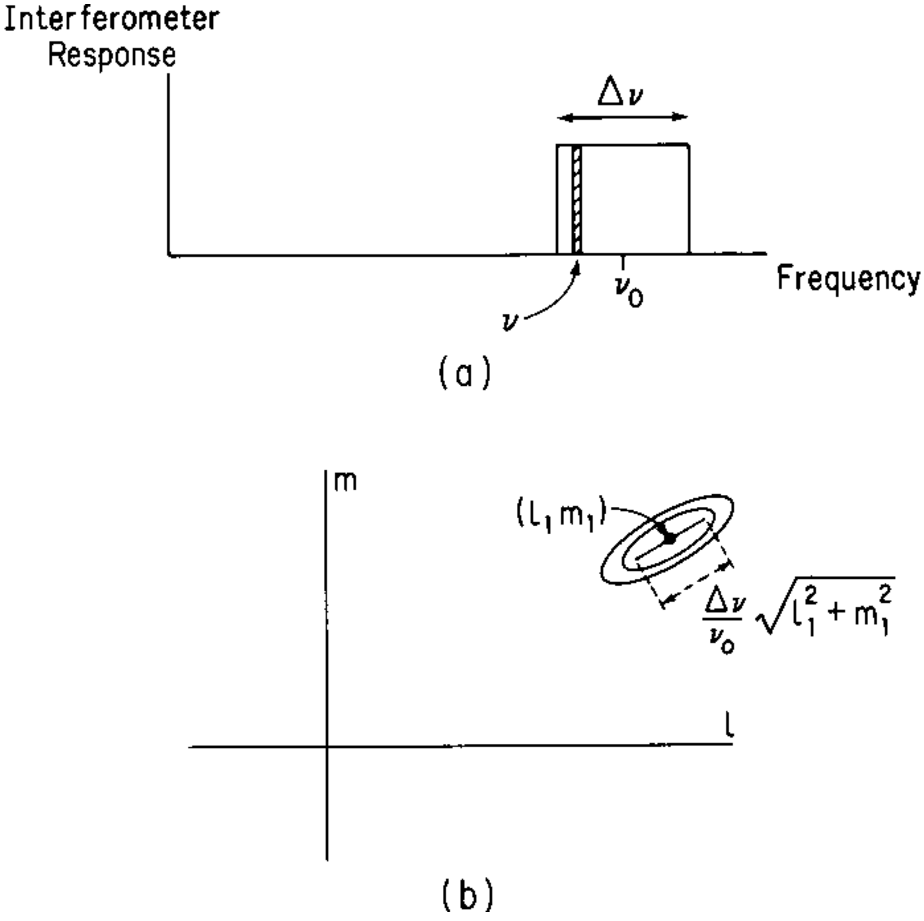


Figure 2–17. (**a**) Idealized rectangular response showing center frequency ν_0 and a narrow band at frequency ν. (**b**) The radial smearing of a point source at (l_1, m_1) in the synthesized image.

brightness I to the synthesized image which is related to the corresponding visibility by

$$V(u,v) \rightleftharpoons I(l,m)\,, \tag{2–34}$$

where the symbol $\rightleftharpoons$ indicates that the two functions constitute a Fourier transform pair, and we have here omitted the functions $\mathcal{A}(l,m)$ and $1/\sqrt{1-l^2-m^2}$ which are usually close to unity. Note that the processes of correlation and Fourier transformation are linear, and that they allow us to consider the synthesized image as the sum of a series of contributions from different parts of the frequency passband. In the derivation of the radio image we assign to V values u_0 and v_0 which are the true values multiplied by ν_0/ν (Eq. 2–33). The effect in the image can be obtained from the similarity theorem of Fourier transforms (e.g., Bracewell 1978), using which one can write

$$V\left(\frac{\nu_0}{\nu}u, \frac{\nu_0}{\nu}v\right) \rightleftharpoons \left(\frac{\nu}{\nu_0}\right)^2 I\left(\frac{\nu}{\nu_0}l, \frac{\nu}{\nu_0}m\right)\,. \tag{2–35}$$

The coordinates of the brightness function are multiplied by the reciprocal of the factor by which the visibility coordinates are multiplied, and a factor $(\nu/\nu_0)^2$ appears in the amplitude to conserve the total integrated brightness. One can envision the effect in the synthesis procedure, in which the data over the full receiving bandwidth $\Delta\nu$ are combined together, as the averaging of a series of images of the same sky brightness distribution, each with a slightly different scale factor and aligned at the (l, m) origin. The range of variation of the scale factor is equal to the variation of ν/ν_0 over the receiving bandwidth. The result of such averaging is clearly to introduce a radial smearing into the brightness distribution, as shown in Figure 2–17b. The angular extent of the smearing at a radial distance $\sqrt{l^2 + m^2}$ from the origin is approximately equal to $\frac{\Delta\nu}{\nu_0}\sqrt{l^2 + m^2}$, and the effect becomes important at distances for which the smearing is comparable with the synthesized beamwidth. An alternative method of imaging with a wide bandwidth is by using a multi-channel receiving system, in which the passband is divided into n frequency channels of width $\Delta\nu/n$. Separate correlators are used for each frequency channel, so the visibility values for each one can be associated with the values of u and v corresponding to the center frequency of the channel. Such systems are also used for spectral line observations. In the (u, v) plane, the elliptical track that represents the projected spacing for any pair of antennas is replaced by a series of n parallel tracks. In effect, the overall transfer function is the sum of n single-channel functions, each scaled in u and v in proportion to the corresponding center frequency of the receiving channel. The sum of the corresponding images shows no radial smearing (we assume that the smearing corresponding to the channel bandwidth $\Delta\nu/n$ is negligible), but since the angular scale of the synthesized beam (point spread function) varies from one channel to the next, the effect of averaging the beam profiles is to reduce unwanted sidelobes. Thus the use of a multi-channel system is a desirable technique in broadband image synthesis, but it requires a significant increase in computing to accommodate n times as many visibility data as in the corresponding continuum observation (see Lecture 21).

12. The Effect of Visibility Averaging

The time averaging of the visibility data at the correlator results in another form of smearing of the image. The data from each correlator are separated into consecutive time intervals of length τ_a, as shown in Figure 2–18a, and only the average value for each interval is retained. In the subsequent processing the averaged visibility samples are assigned (u, v) values corresponding to the mid-points of the averaging intervals, although the observed data extend over a range $\pm\tau_a/2$ relative to each such instant. The effect in the synthesized image can be most easily explained for an observation of a source at the celestial pole. The (u, v) plane is then normal to the Earth's axis, and the transfer function consists of a series of circles, concentric about the u-v origin, as in Figure 2–18b. Each circle is generated by a spacing vector rotating with angular velocity ω_e equal to that of the Earth. Thus a time offset τ in the assignment of (u, v) values results in a rotation of the visibility function about the (u, v) origin through an angle $\omega_e\tau$. In the Fourier transformation, such a rotation results in an equal rotation of the image. Thus the effect of the time averaging can be envisioned

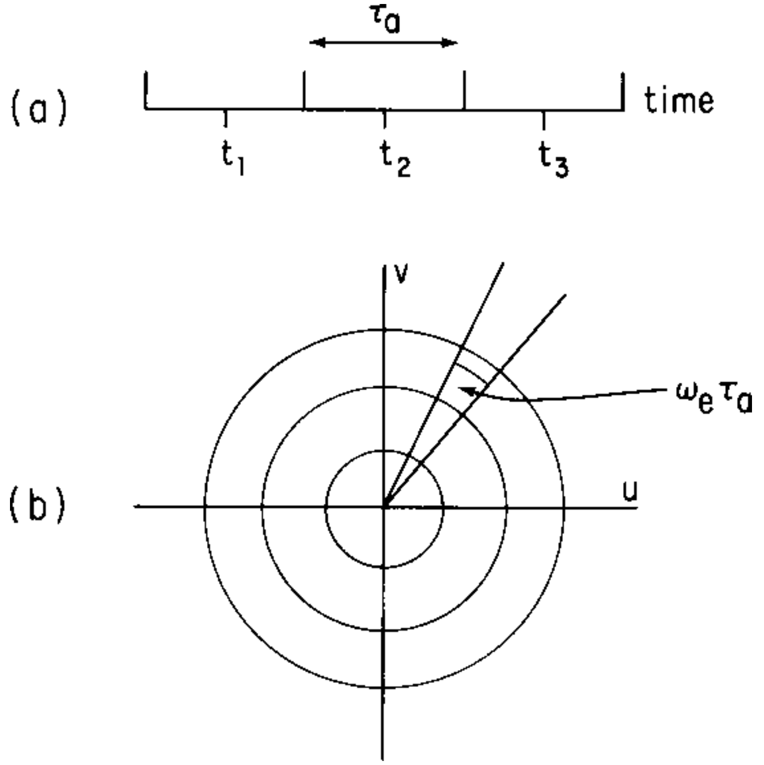


Figure 2–18. **(a)** Consecutive time intervals of duration τ_a over which the visibility is averaged. **(b)** Circular loci in the (u, v) plane which result from the continuous observation of a source close to the celestial pole. In a time interval τ_a, the baseline vectors which generate the loci move through an angle $\omega_e \tau_a$.

as an averaging of a series of images that are aligned at the (l, m) origin, but have angular offsets distributed over a range $\pm\omega_e\tau_a/2$. At a point (l, m) the extent of the smearing is approximately $\omega_e\tau_a\sqrt{l^2 + m^2}$. The direction of the smearing is orthogonal to that resulting from the bandwidth effect, and the two effects are of equal magnitude if $\Delta\nu/\nu_0 = \omega_e\tau_a$.

For a source at a lower declination the curves in the transfer function become ellipses, and are centered at the (u, v) origin only for East–West baselines. In this latter case the expansion of the v-axis by a factor $\csc\delta$ restores the circularity, so in an image plane in which the m-axis (North–South) is compressed by a factor $\sin\delta$, the effect is again one of circumferential smearing. In the general case of a non-polar source and non- East–West baselines, the effect of time averaging cannot be described simply in terms of a rotational smearing.

References

Arsac, J. 1955, *C. R. Acad. Sci.*, 240, 942–945.

Bracewell, R. N. 1966, Report on the Fifteenth General Assembly of URSI, Pub. 1468, National Academy of Sciences (Washington, D.C.), pp. 243–244.

Bracewell, R. N. 1978, *The Fourier Transform and its Applications*, Second Edition, McGraw–Hill, New York.

Bracewell, R. N., Colvin, R. S., D'Addario, L. R., Grebenkemper, C. J., Price, K. M., & Thompson, A. R. 1973, *Proc. IEEE*, 9, 1249–1257.

Christiansen, W. N. & Högbom, J. A. 1985, *Radiotelescopes*, Second Edition, Cambridge University Press (Cambridge, England).

Fomalont, E. B. 1973, *Proc. IEEE*, 61, 1211–1218.

Fomalont, E. B. & Wright, M. C. H. 1974, in *Galactic and Extragalactic Radio Astronomy*, First Edition, G. L. Verschuur & K. I. Kellermann, Eds., Springer–Verlag, New York, pp. 256–290.

Frater, R. H. 1984, *Proc. Astron. Soc. Australia*, 5, 440–445.

Granlund, J., Thompson, A. R., & Clark, B. G. 1978, *IEEE Trans. Electromag. Compat.*, EMC-20, 451–453.

Högbom, J. A. & Brouw, W. N. 1974, *A&A*, 33, 289–301.

Meeks, M. L., Ed. 1976, *Methods of Experimental Physics*, Vol. 12C, Academic Press, New York; see chapters on interferometry.

Michelson, A. A. 1890, *Phil. Mag.*, Ser. 5, 30, 1–21.

Moffet, A. T. 1968, *IEEE Trans. Antennas Propagat.*, AP–16, 172–175.

Napier, P. J., Thompson, A. R., & Ekers, R. D. 1983, *Proc. IEEE*, 71, 1295–1320.

Napier, P. J., Bagri, D. S., Clark, B. G., Rogers, A. E. E., Romney, J. D., Thompson, A. R., & Walker, R. C. 1994, *Proc. IEEE*, 82, 658–672.

Preston, R. A., Burke, B. F., Doxsey, R., Jordan, J. F., Morgan, S. H., Roberts, D. H., & Shapiro, I. I. 1983, in *Very Long Baseline Interferometry Techniques*, F. Biraud, Ed., Cepadues (Toulouse, France), pp. 417–431.

Rohlfs, K. 1986, *Tools of Radio Astronomy*, Springer–Verlag, Berlin.

Ryle, M. 1952, *Proc. Roy. Soc.*, 211A, 351–375.

Ryle, M. 1962, *Nature*, 194, 517–518.

Ryle, M. 1972, *Nature*, 239, 435–438.

Swenson, G. W., Jr. & Mathur, N. C. 1968, *Proc. IEEE*, 56, 2114–2130.

Thompson, A. R., Clark, B. G., Wade, C. M., & Napier, P. J. 1980, *ApJS*, 44, 151–167.

Thompson, A. R. & D'Addario, L. R. 1982, *Radio Science*, 17, 357–369.

Thompson, A. R., Moran, J. M., & Swenson, G. W., Jr. 1986, *Interferometry and Synthesis in Radio Astronomy*, John Wiley & Sons, New York.

Walker, R. C. 1984, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 53–65.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

3. The Primary Antenna Elements

P. J. Napier
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. The primary antenna elements are one of the most important pieces of equipment in a synthesis telescope. Because the performance properties of the antennas can affect the quality of the synthesized images in a number of fundamental ways, the relevant antenna design and performance parameters are reviewed in this lecture.

1. Introduction

In this lecture I describe the design and performance of the primary antenna elements which are used to sample the electromagnetic field radiated by the observed radio source. In general, throughout these lectures, few details are given of the equipment that is used in synthesis arrays. It is appropriate, however, to treat the antenna elements more completely because of the many ways in which they can directly affect the quality of the images produced by the array. The important properties of the primary antenna that can affect the image include aperture efficiency, pointing accuracy, beam circularity, sidelobe level, polarization purity and noise temperature. The techniques needed to achieve good antenna performance with respect to these parameters can be found in the modern antenna engineering literature (see, for example, Rudge et al. 1982; Kraus 1988; Olver et al. 1994; Balanis 1997) so I will not give details in this lecture. Instead, I shall emphasize why these parameters are important.

Figure 3–1 shows a simple block diagram of the major pieces of equipment required in a synthesis telescope. For the purposes of this lecture I will define the primary antenna element to be the piece of equipment that intercepts the propagating electromagnetic wave from the observed source and makes a sample of it available at the input to the first low-noise amplifier, either as an electric current on a cable or as a field in a single-mode waveguide. Thus, for reflector antennas, I include the feed and its polarization splitter as part of the antenna. At the output of the antenna the signal is at the radio, or sky, frequency $\nu_{\rm RF}$. As shown in Figure 3–1, the signal undergoes various frequency translations as it propagates through the electronics system. In this lecture I will not be concerned with any of the equipment after the antenna. The correlator is treated in detail in Lecture 4, and a discussion of the importance of the noise temperature of the receiver is given in Lecture 9. In general the receiver, intermediate frequency, transmission line, local oscillator and baseband portions of the electronics system all have the requirement of good amplitude and phase stability. These requirements and others such as bandpass shape control, low spurious signal generation and good signal isolation are discussed in references such as Napier, Thompson & Ekers (1983) and Thompson, Moran & Swenson (1986).

2. Basic Antenna Formulas

In this section I introduce a few standard antenna formulas which will be useful in understanding how the properties of the primary antenna element affect a

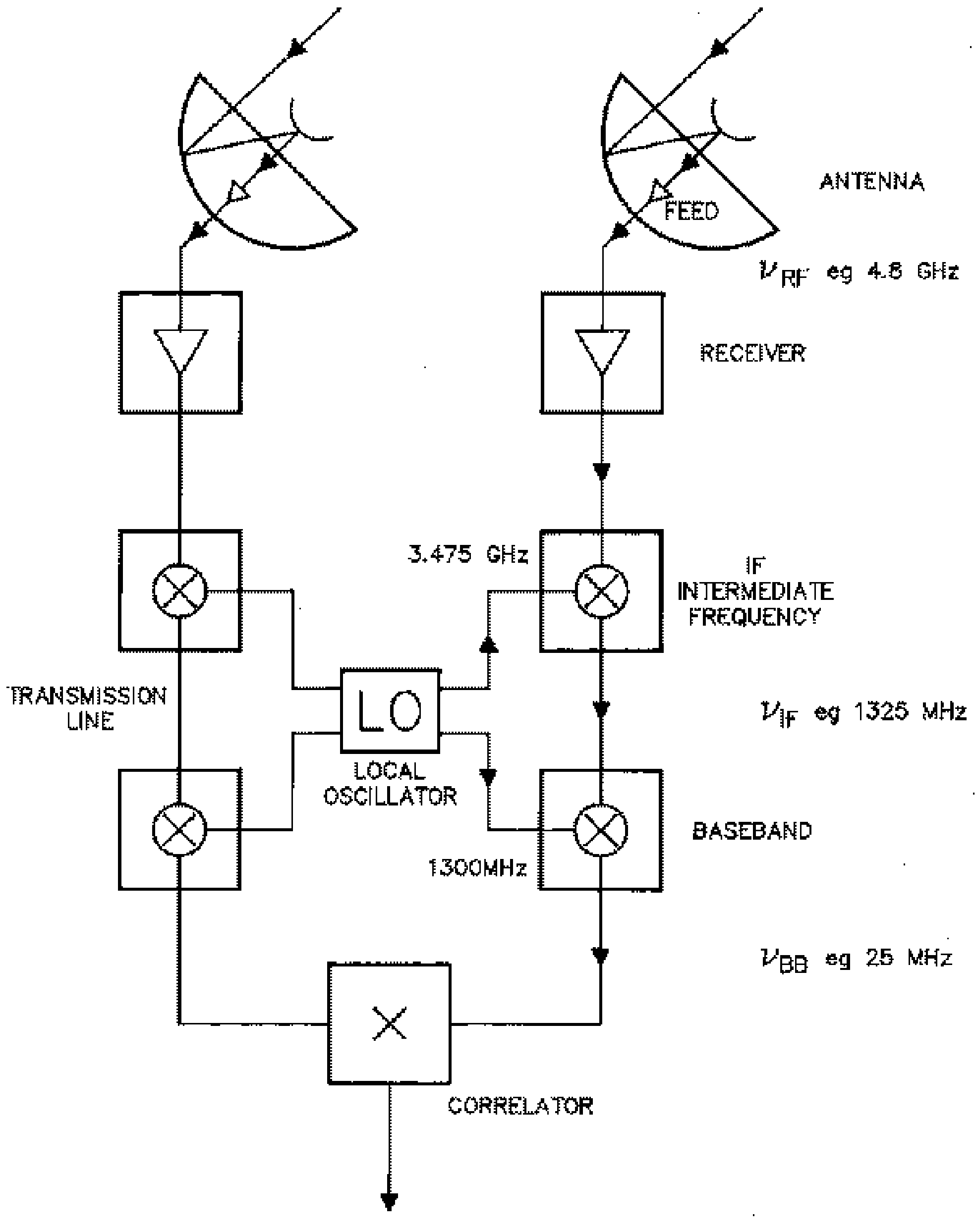


Figure 3–1. A simplified block diagram of the electronic equipment used to produce the correlation from one antenna pair in a synthesis telescope. The signal frequencies given as examples at various points through the electronics chain are typical of the VLA observing at 4.8 GHz.

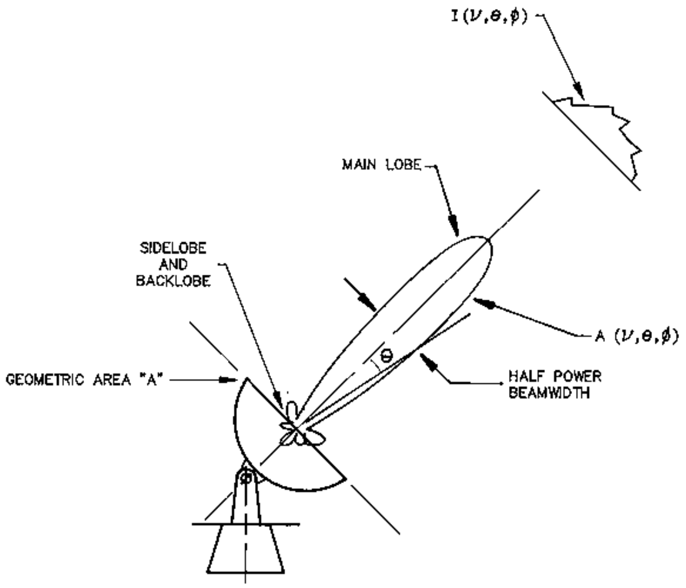


Figure 3–2. The reception pattern of an antenna.

synthesized image. Derivation of these expressions can be found in standard textbooks (e.g., Kraus 1986, Chapter 3).

We require the concept of the effective collecting area of the primary antenna $A(\nu, \theta, \phi)$ (in units of m^2), where ν is frequency and θ and ϕ are direction coordinates. If the antenna is pointed at a source with brightness $I(\nu, \theta, \phi)$ $\mathrm{W\,m^{-2}\,Hz^{-1}\,sr^{-1}}$ (see Figure 3–2), then the power P (in watts) received by the antenna in bandwidth $\Delta\nu$ from element $\Delta\Omega$ of solid angle is given by

$$P = A(\nu, \theta, \phi) I(\nu, \theta, \phi) \Delta\nu \Delta\Omega \; . \tag{3–1}$$

The normalized antenna reception pattern $\mathcal{A}$, or power pattern as it is often called, is defined as

$$\mathcal{A}(\nu, \theta, \phi) = A(\nu, \theta, \phi)/A_0 \; , \tag{3–2}$$

where A_0 (m^2) is the response at the center of the main lobe of $A(\nu, \theta, \phi)$ and is called the effective area of the antenna. The beam solid angle, Ω_A, of the power pattern is defined as

$$\Omega_A = \iint_{\text{all sky}} \mathcal{A}(\theta, \phi) \, d\Omega \, . \tag{3–3}$$

An important fundamental relationship in antenna theory states that the product of the effective area and the beam solid angle is equal to the square of the wavelength (Kraus 1986, page 6–5),

$$A_0 \Omega_A = \lambda^2 \, . \tag{3–4}$$

Ω_A is a measure of the field of view of the synthesis telescope. If $\mathcal{A}$ is everywhere equal to 1, then Ω_A has its maximum possible value of 4π and the primary antenna is isotropic and can see the whole sky with equal sensitivity. In this case the synthesis telescope could in principle make an image of the whole sky all at once. A large field of view is desirable, but Eq. 3–4 shows that, for any given frequency, when Ω_A is a maximum, A_0 is a minimum and so the power received is also a minimum. This means that the sensitivity is at a minimum. As the collecting area is increased to improve sensitivity, Eq. 3–4 dictates that the field of view necessarily decreases. This tradeoff between field of view and sensitivity has to be made when selecting the size of the primary antenna elements for any synthesis telescope. It requires consideration of the size/cost curve for the antennas and the expected size and spectral index properties of the radio sources to be observed.

For antennas that have a well defined physical collecting area, such as reflector, lens or horn antennas, the ratio of A_0 to the physical area A of the aperture is called the aperture efficiency η, a dimensionless quantity less than unity:

$$A_0 = \eta A \,. \tag{3–5}$$

The final antenna relationship that we will find useful is the Fourier transform relationship between the complex voltage distribution of the field, $f(u,v)$,[1] in the aperture of an antenna and the complex far-field voltage radiation pattern, $F(l,m)$, of the antenna (Kraus 1986, Section 6–8):

$$F(l,m) = \iint_{\text{aperture}} f(u,v) e^{2\pi i(ul+vm)}\, du\, dv \,, \tag{3–6}$$

and

$$f(u,v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(l,m) e^{-2\pi i(ul+vm)}\, dl\, dm \,. \tag{3–7}$$

Figure 3–3 shows the form of $f(u,v)$ and $F(l,m)$. The radiation pattern coordinates are given by

$$u = \sin\theta\cos\phi \quad \text{and} \quad v = \sin\theta\sin\phi \,. \tag{3–8}$$

The voltage and power patterns are related by $\mathcal{A} = |F|^2$. The Fourier transform relationship between an antenna aperture distribution and its radiation pattern is analogous to the one between the source brightness distribution and its visibility function. In the analogy, aperture distribution and brightness distribution are the analogous quantities because they both are of compact support,[2] although the analogy is not perfect because the aperture distribution is complex whilst brightness is real. Both antenna radiation pattern and source visibility function have the same general form of a main central lobe surrounded by lesser sidelobes, but the radiation pattern does not have to be Hermitian (i.e., need not be conjugate symmetric).

[1] We have chosen u and v rather than the more traditional x and y to represent the spatial coordinates of the antenna in order to be consistent with the similar use of u and v to represent spatial coordinates in the synthesised aperture throughout this book.

[2] This is not strictly so in the case of the brightness distribution, but it is essentially so.

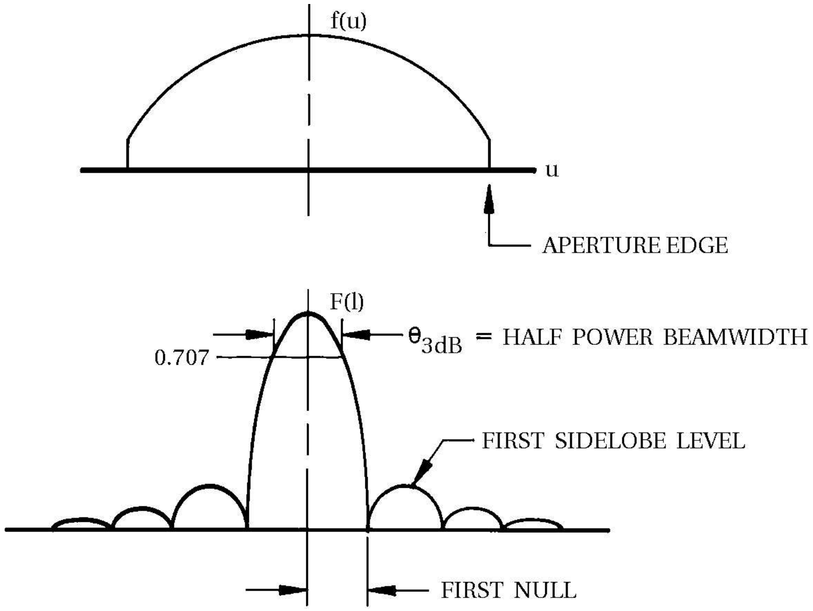


Figure 3–3. The Fourier transform relationship between an antenna aperture distribution and its far-field radiation pattern. The form of the aperture distribution, $f(u)$, and the radiation pattern, $F(l)$, are shown for a one-dimensional example. In general both quantities are complex. Only the amplitudes are shown here.

The form of $f(u,v)$ for an antenna is determined by the way in which the antenna feed illuminates the aperture. In general, the more that $f(u,v)$ is tapered at the edge of the aperture, the lower will be the aperture efficiency and the sidelobes and the broader the main beam. Tabulations of a wide variety of $f(u,v)$, and their $F(l,m)$, can be found in antenna textbooks (Hansen 1964, p. 66; Rudge *et al.* 1982, Table 1.2). For example, the VLA antennas are designed to have uniform illumination ($f(u,v) =$ constant) over the whole aperture, except where the aperture is blocked by the subreflector and its support struts. In this case, for a circularly symmetric aperture of diameter D, $F(u) = J_1(\pi Du)/u$, which has the properties: first sidelobe level $= -17.6$ dB, half power beamwidth $= 1.02\lambda/D$, and position of first null $= 1.22\lambda/D$. These are in good agreement with measured beam parameters for the VLA 25-meter diameter reflector, except for the first sidelobe level, which is modified by aperture blockage as shown in Figure 3–8.

3. General Antenna Types

In this section I will discuss some of the general considerations that determine the selection of a primary antenna element for a synthesis array.

3.1. Wavelength range

The most important factor governing the selection of the primary antenna element is the frequency range to be observed by the synthesis telescope. Typically,

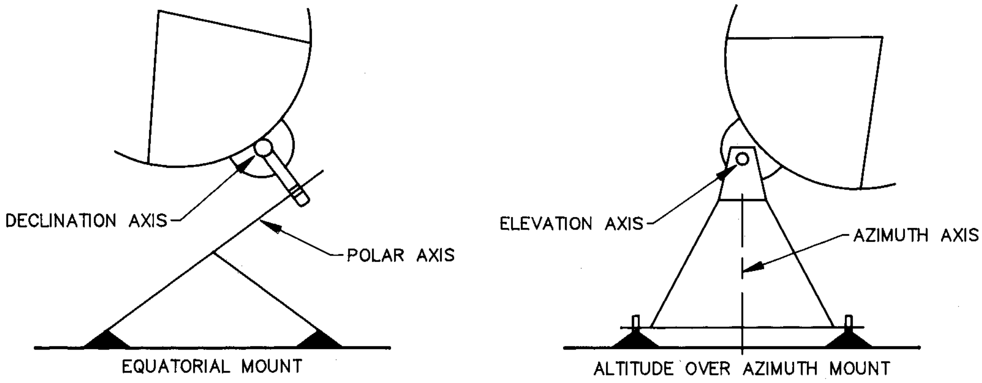


Figure 3–4. Equatorial and Altazimuth antenna mounts.

for wavelengths longer than about 1 m, wire antennas are used. These include dipoles, yagis, spirals and helices. Sometimes small arrays of these elements are used to increase collecting area. For wavelengths shorter than approximately 1 m, reflector antennas are common, and occasionally horn antennas are used for very wide field of view instruments. The changeover frequency from wire to reflector antennas is not sharply defined, and, around 1-m wavelength, combinations of the two types can be found. Thus the Molonglo Synthesis Telescope at 800 MHz uses a cylindrical paraboloidal reflector with a reflective surface formed with parallel wires, and the VLA at 75 MHz uses a solid 25-m diameter reflector fed by a simple wire crossed dipole. References for the different antenna elements that have found application in synthesis telescopes can be found in the comprehensive listing of synthesis arrays given in Napier, Thompson & Ekers (1983).

Since wire antennas are simpler and less expensive to build than reflector antennas one might ask why they are not used at all wavelengths. The answer is contained in Equation 3–4. As an antenna design is scaled from one wavelength to another, so that its pattern $\mathcal{A}$ is the same, its collecting area increases as λ^2. Thus, whilst wire antennas have sufficient collecting area to be useful at long wavelengths, they are too small to be primary elements at short wavelengths. Since most of the currently active synthesis telescopes use reflector antennas I will concentrate on them for the rest of this lecture.

3.2. Types of reflector antennas

The two major choices that have to be made when designing a reflector antenna for use in a synthesis array are the choice of mount and the choice of optics.

Choice of mount. The choice of the mount is usually between an equatorial mount and an elevation-over-azimuth (altazimuth) mount, as shown in Figure 3–4. The elevation-over-azimuth mount has the advantage of simplicity and hence lower cost. Gravity always acts on the reflector in the same plane, and this eases the problem of designing to keep the reflector profile accurate as the antenna tracks an astronomical source. The major disadvantage of this mount is that, as the antenna tracks, the aperture rotates with respect to the source (Thompson,

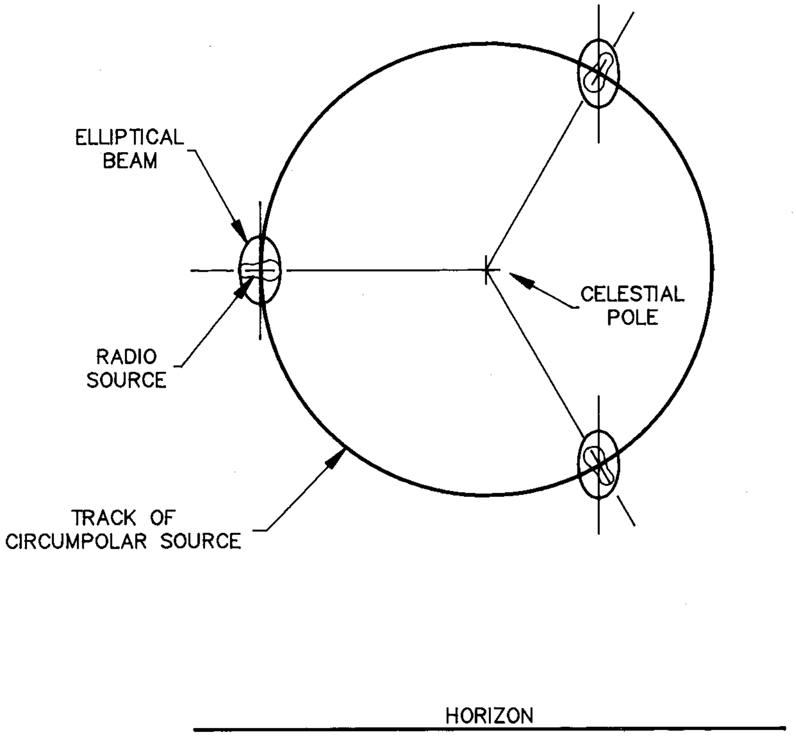


Figure 3–5. Demonstration of the rotation of the beam of an altazimuth antenna with respect to an astronomical source. A circumpolar double-lobed source, extended in the North–South direction, is shown at three different hour angles. The antenna beam is extended in the vertical direction.

Moran & Swenson 1986, p. 97). This rotation is about the line from the center of the aperture to the source and means that the antenna beam rotates with respect to the source. If the source size is of the order of the beam size, and if the beam is not circularly symmetric, this rotation will cause the apparent brightness distribution to vary. Figure 3–5 gives a simple demonstration of this effect. Since aperture blockage usually makes the beam sidelobe pattern non-circularly symmetric, and the antenna instrumental polarization is not circularly symmetric, the dynamic range of total intensity images of very large sources and polarization images of extended sources will be limited by this effect. Observers of extended sources need to consider this effect when judging the fidelity of subtle features in the images of these sources. Another minor disadvantage of this mount is that a source passing close to the zenith of the antenna cannot be tracked near the zenith because of the high rates of azimuth rotation needed.

The equatorial mount has the advantage that, since the polar axis is aligned parallel to the axis of rotation of the Earth, rotation about this single axis is adequate to track an astronomical object moving at sidereal rate. Its principal advantage is that it does not suffer from the beam rotation problem discussed above for the alt–az mount. The principal disadvantage is that as the reflector tracks, gravity does not act always in the same plane. This, together with the inclined polar axis, significantly increases the complexity of the design, with a resulting increased cost. A minor disadvantage of this mount is that it cannot observe sources close to the horizon in the direction away from the celestial pole.

A final important point is that, irrespective of what type of mount is chosen, it is important that antenna structures be made as identical as possible between different elements of the array. This will minimize the effect of many kinds of deformations and other errors, or at least make it easier to calibrate them.

Choice of optics. There are a number of different optical systems that can be used to feed a large radio reflector (Rudge et al. 1982, Section 3.6). Figure 3–6 shows some of the feed systems that have been used for radio telescope reflectors. The *prime focus system* (as in the Westerbork Synthesis Telescope) has the advantage that it can be used over the full frequency range of the reflector, including the lowest frequencies where secondary focus feeds become impractically large. The disadvantages of the prime focus are that space for, and access to, the feed and receiver is restricted and spillover noise from the ground decreases sensitivity. All of the *multiple reflector systems* (Figure 3–6(b)–(f)) have the advantage of more space, easier access to the feed and receiver, and reduced noise pickup from the ground. In addition, the primary and secondary reflectors can be shaped to provide more uniform illumination in the main reflector aperture, as described in Section 4.1.

The *off-axis Cassegrain* (e.g., in the VLA and VLBA) is particularly suitable for synthesis telescopes needing frequency flexibility, because many feeds can be located in a circle around the main reflector axis so that changing frequency simply requires a rotation of the subreflector around this axis. The disadvantage of this geometry is that the asymmetry degrades polarization performance. The *Naysmith geometry* (e.g., the Owens Valley Millimeter Array) provides a receiver cabin external to the antenna structure, whilst the *beam waveguide feed* (e.g., the Nobeyama 45-m millimeter wavelength telescope) provides maximum convenience by locating the feeds and receivers at ground level. The *offset Cassegrain* (e.g., the Bell Labs millimeter telescope) has no blockage and so can have a circularly symmetric beam with low sidelobes. This makes it an attractive choice for wide field-of-view synthesis telescopes, but the increased complexity of reflector panel tooling and subreflector support structure leads to increased cost.

4. Antenna Performance Parameters

In this section I discuss some of the performance parameters of primary antenna elements that can directly affect the quality of images made with synthesis telescopes.

4.1. Aperture efficiency

The antenna aperture efficiency, defined in Equation 3–5, directly impacts the sensitivity of the synthesis telescope (Lecture 9). The aperture efficiency, η, is the product of a number of different loss factors,

$$\eta = \eta_{\mathrm{sf}}\,\eta_{\mathrm{bl}}\,\eta_{\mathrm{s}}\,\eta_{\mathrm{t}}\,\eta_{\mathrm{misc}}\,, \qquad (3\text{–}9)$$

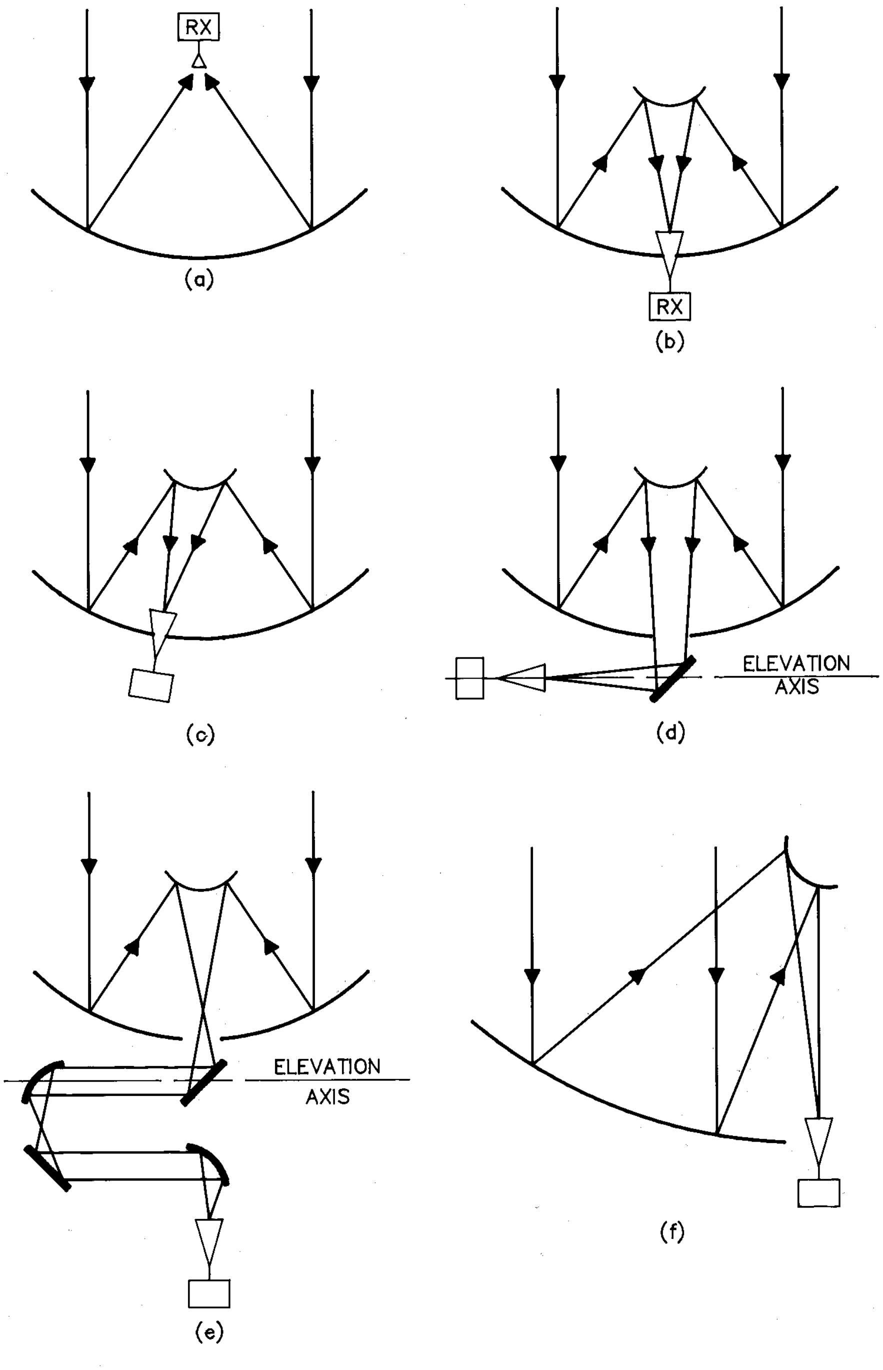


Figure 3–6. Optical systems for radio telescope reflectors. **(a)** Prime focus, **(b)** Cassegrain, **(c)** Off-axis Cassegrain, **(d)** Naysmith, **(e)** Beam waveguide, **(f)** Offset Cassegrain.

where $\eta_{\rm sf}$ = reflector surface efficiency, $\eta_{\rm bl}$ = reflector blockage efficiency, $\eta_{\rm s}$ = feed spillover efficiency, $\eta_{\rm t}$ = illumination taper efficiency, and $\eta_{\rm misc}$ = miscellaneous efficiency losses due to reflector diffraction, feed position phase errors, and feed match and loss.

Surface efficiency. This factor accounts for aperture efficiency loss due to inaccuracies in the reflector profile. If the reflector has errors, then the electric field from different parts of the aperture will not add together perfectly in phase at the feed, leading to a decrease in received power. Ruze (1966) gives an expression for surface efficiency

$$\eta_{\rm sf} = e^{-(4\pi\sigma/\lambda)^2} , \tag{3–10}$$

where σ is the r.m.s. surface error, with the errors assumed to be random from a Gaussian population and uncorrelated from one point to another in the aperture. In a Cassegrain system σ is an appropriately defined composite r.m.s. error of the primary and secondary reflector surfaces. If the errors are correlated over significant fractions of the aperture, then additional terms are required on the right hand side of Eq. 3–10 (Ruze 1966). Eq. 3–10 predicts that for an r.m.s. error of $\lambda/16$, $\eta_{\rm sf} = 0.54$, which is often taken as the useful upper frequency limit for a reflector. As well as the loss of sensitivity resulting from a low value of $\eta_{\rm sf}$, one has to be concerned with the quality of the primary beam. Ruze (1966) shows that the surface errors produce a broad scatter pattern that surrounds the main lobe of the beam and can be higher than the usual diffraction-limited sidelobes. This scatter pattern could enhance image artifacts caused by sources near the primary beam. For a reflector of diameter D, if the reflector errors are correlated over distances D/N then the scatter pattern will be N times broader than the diffraction-limited main lobe. Figure 3–7 shows the scatter pattern for surface errors of size $\lambda/16$ r.m.s. and $N = 10$, as might be the case, for example, if the errors are due to incorrect alignment of otherwise accurate reflector panels.

Good $\eta_{\rm sf}$ performance requires careful structural design for wind, thermal and gravitational loading, together with precise reflector panels and an accurate panel setting technique.

Aperture blockage. The feed or subreflector and its multi-legged support structure block the aperture of a reflector antenna, as shown in Figure 3–8. This typically results in a blockage efficiency in the range $0.75 < \eta_{\rm bl} < 0.95$. $\eta_{\rm bl}$ is given (Rudge et al. 1982, p. 179) by

$$\eta_{\rm bl} = \left(1 - \frac{\text{effective blocked area}}{\text{total area}}\right)^2 . \tag{3–11}$$

The effective blocked area is the blocked area weighted for the illumination taper in the aperture. Similarly, the total area is weighted for the illumination taper in the aperture. Equation 3–11 shows, for small blockage, that the loss in efficiency is twice the fractional blocked area. As well as the loss in aperture efficiency, the increase in antenna beam sidelobe level due to blockage is important for synthesis telescopes. Using the Fourier transform relationship of Eq. 3–6, the form of the antenna voltage pattern with blockage can be calculated as the unblocked voltage pattern minus the voltage patterns of the blocked areas. This

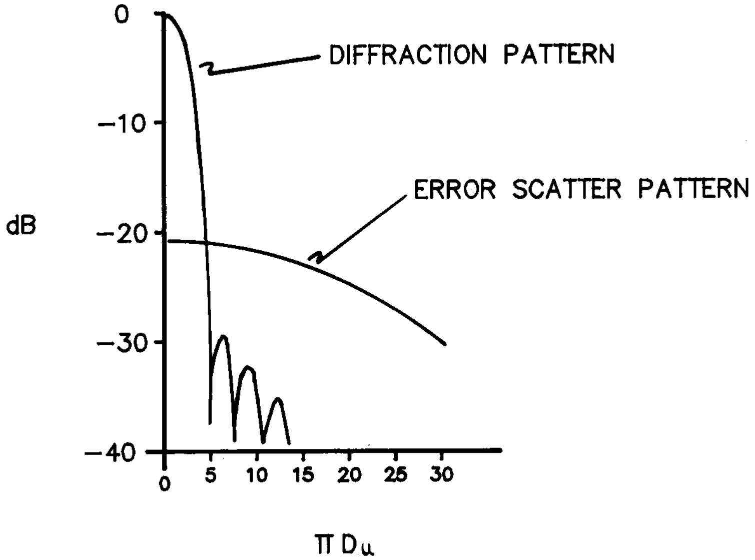


Figure 3–7. Diffraction pattern and surface error scatter pattern for r.m.s. surface errors $\lambda/16$ correlated over distances $D/10$. The aperture has a −12 dB edge taper. The total pattern is the power sum of the diffraction and error patterns. [From Ruze 1966.]

is shown in Figure 3–8, where the typical pattern of enhanced sidelobes in planes orthogonal to the subreflector support legs can be seen.

Feed spillover efficiency. This effect can most easily be understood by considering the antenna in transmission, rather than reception mode. The feed spillover efficiency is the fraction of the power radiated by the feed that is intercepted by the subreflector for a Cassegrain feed, or by the main reflector for a prime focus system. Clearly, power that does not intercept the reflector is lost, and we can be confident that a similar loss occurs in reception mode by invoking the Reciprocity Principle (Rudge et al. 1982, p. 11). η_s is given (Rudge et al. 1982, p. 170) by

$$\eta_s = \frac{\int_0^{2\pi}\int_0^{\theta_R} P_f(\theta,\phi)\sin\theta\,d\theta\,d\phi}{\int_0^{2\pi}\int_0^{\pi} P_f(\theta,\phi)\sin\theta\,d\theta\,d\phi}\,, \tag{3–12}$$

where $P_f(\theta,\phi)$ is the power pattern of the feed and θ_R is the angle subtended by the reflector (see Figure 3-9). η_s increases as $P_f(\theta_R)$ is reduced, but η_t decreases, so a tradeoff between spillover and illumination efficiency must be made. The shaped reflector systems discussed in the next section ease this problem. Typically, $0.70 < \eta_s < 0.97$, with the higher values requiring shaped reflectors. As well as the loss of sensitivity, the impact of spillover on the sidelobes of the antenna must be kept in mind. At θ_R, the gain of the feed may be well above the diffraction sidelobes of the primary aperture, causing higher than expected antenna sidelobes at distance θ_R away from the peak of the main beam. This problem is worst at the lowest frequency, where the gain of the primary aperture

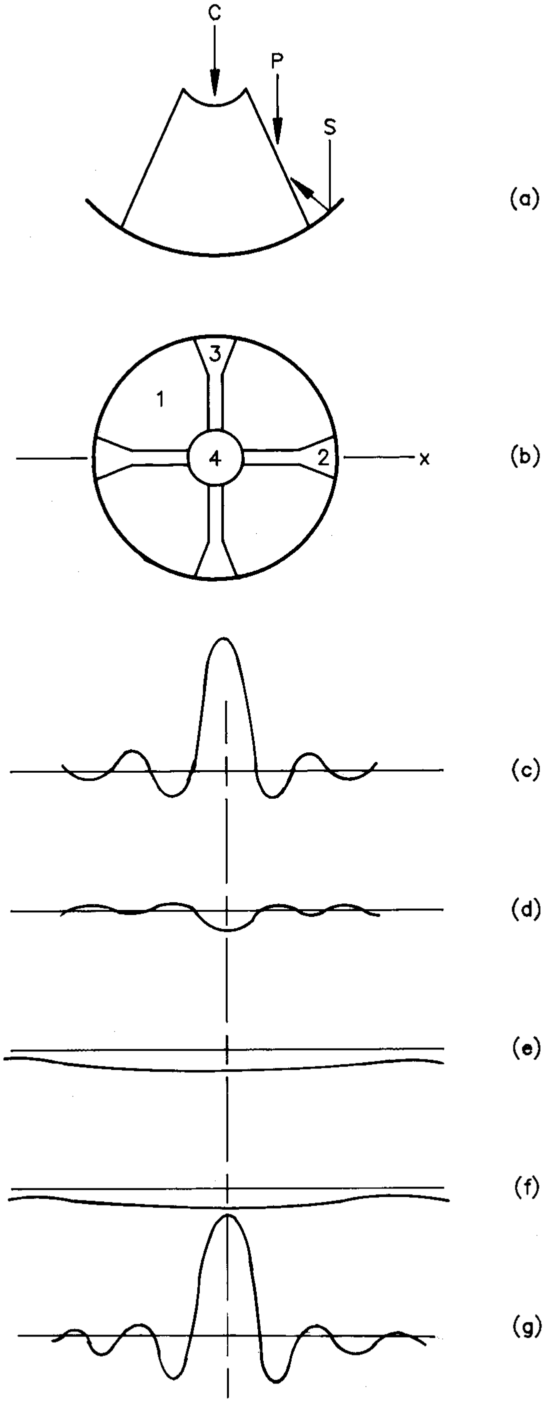


Figure 3–8. Effect of aperture blockage. **(a)** The three kinds of blockage in a reflector. c is central blockage, p is plane wave blockage on the struts, and s is spherical wave blockage on the struts. **(b)** The resulting aperture blockage for a quadrupod subreflector support. **(c)** Unblocked pattern in x-plane. **(d)** Blocked pattern for area 2. **(e)** Blocked pattern for area 3. **(f)** Blocked pattern for area 4. **(g)** Total pattern with blockage $= c + d + e + f$.

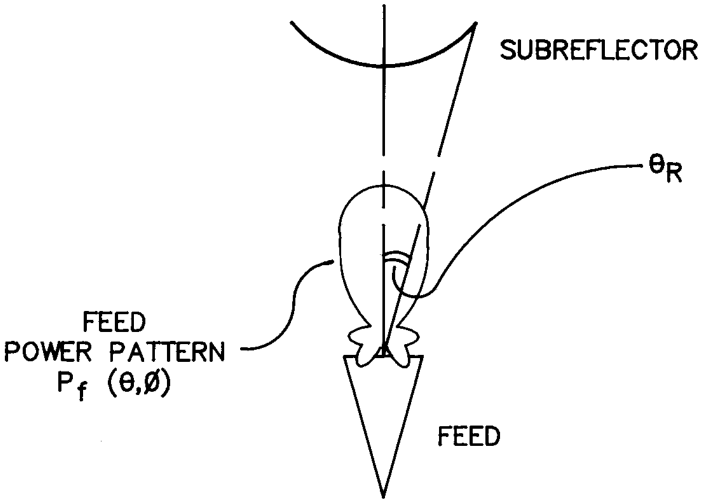


Figure 3–9. Feed spillover.

is lowest. For example, for the VLA at $\lambda = 21$ cm, the feed spillover past the subreflector is about 20 dB higher than the diffraction-limited sidelobes at a distance $\theta_R = 9°$ off the peak of the beam. This effect can enhance the effect of confusing sources outside the expected field of view of the synthesis image.

Illumination taper efficiency. Illumination taper efficiency accounts for the loss in collecting area due to the fact that the feed pattern usually illuminates the outer parts of the primary reflector at a lower level than the inner part. η_t is given (Rudge et al. 1982, p. 171) by

$$\eta_t = \frac{\left(\iint_{\text{aperture}} |f(u,v)|\, du\, dv\right)^2}{A \iint_{\text{aperture}} |f(u,v)|^2\, du\, dv}\,. \tag{3–13}$$

η_t equals 1 for uniform illumination $(f(u,v) \equiv 1)$. Typically, to prevent η_s from being too low, an illumination taper of about 10 dB is used at the edge of the aperture, which results in a η_t in the range 0.7 to 0.8. This can be improved significantly if a shaped Cassegrain (Rudge et al. 1982, p. 247) reflector system is used. With this technique, the subreflector is deformed slightly from the usual hyperbolic shape in such a way that the feed pattern, after reflection from the subreflector, is significantly modified, with increased illumination at the edge of the aperture (see Figure 3–10). Since the shape of the subreflector is altered, the main reflector must also be modified slightly from a paraboloid to avoid aperture phase errors. Using this technique, values of η_t close to 1 can be achieved with reflector edge tapers in the range −15 dB to −20 dB. The resulting $\eta_t\eta_s$ product can be 20% to 30% better than with unshaped systems.

Some disadvantages of shaped Cassegrain geometries, which do not usually preclude their use for synthesis telescopes, include increased sidelobes due to the uniform illumination, no prime focus operation above a frequency of about 1 GHz because of the shaped main reflector, and very bad beam degradation if

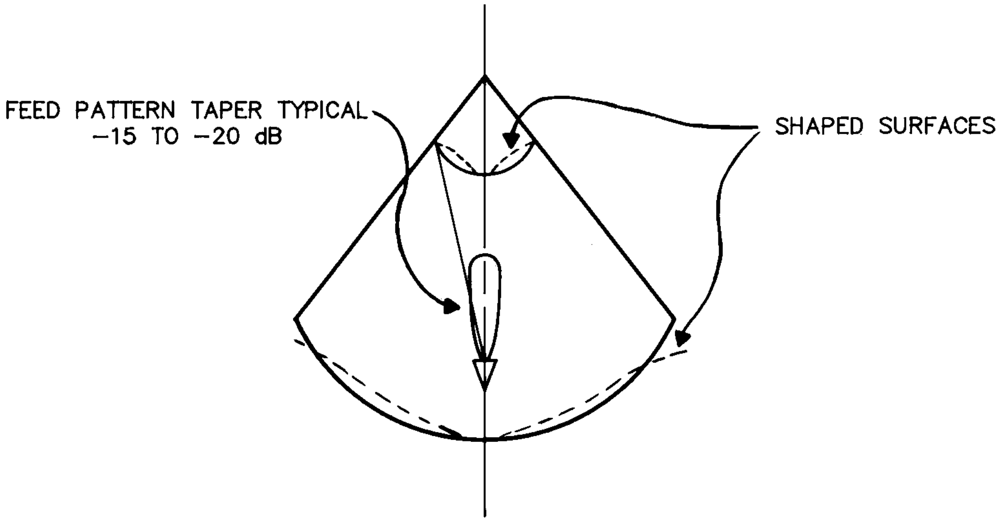


Figure 3–10. Shaped Cassegrain reflector system. The subreflector is deformed to increase illumination at the edge of the main reflector.

the feed is moved away from the secondary focal point. This latter problem may be a limitation for synthesis arrays intended to obtain very wide fields of view by using multiple feeds. An interesting alternative to shaping the reflectors is to place a correctly designed lens in front of the feed (Hudson et al. 1987), which has the advantage of being easily removable if desired.

Example of VLA performance. In this section I will give some aperture efficiency and radiation pattern data for the VLA as typical examples of the performance of the primary elements of a synthesis array. Table 3–1 shows the predicted values for the various factors discussed above for four of the VLA observing wavelengths.

Table 3–1. VLA Performance Examples

λ	$\eta_{\rm sf}$	$\eta_{\rm bl}$	$\eta_{\rm s}$	$\eta_{\rm t}$	$\eta_{\rm diff}$	$\eta_{\rm misc}$	$\eta_{\rm pred}$	$\eta_{\rm meas}$
20 cm	1.0	.85	.82	.98	.86	.94	.55	.51
6 cm	.97	.85	.92	.98	.96	.94	.67	.65
2 cm	.85	.85	.90	.95	.98	.94	.57	.52
1.3 cm	.68	.85	.90	.95	.99	.94	.46	.43

In Table 3–1 $\eta_{\rm diff}$, the diffraction loss of the subreflector, is significant at 20-cm wavelength, where the subreflector is too small; $\eta_{\rm pred}$ is the total predicted efficiency, the product of the six factors to its left; and $\eta_{\rm meas}$ is the actual measured efficiency.

A powerful diagnostic measurement for reflector antennas is known as the "holographic" measurement (Scott & Ryle 1977; Mayer et al. 1983; Kraus 1986, section 6-22) and is ideally suited for use on the primary elements of a synthesis array. This measurement is based on the Fourier transform relationship of Eq. 3–7. $F(l, m)$ is measured by scanning one of the antennas, as in Figure 3–1, back and forth across a strong point source while the other antenna continuously tracks the source. $f(u, v)$, the complex aperture distribution of the scanning antenna, is then computed using Equation 3–7. $|f(u, v)|$ then directly shows the aperture blockage and illumination, whilst $\arg(f(u, v))$ provides information about reflector surface errors and feed location errors. Figure 3–11 shows the results of a holographic measurement of a VLA antenna. $|F(l, m)|$ shows the expected sidelobe enhancement in the planes of the subreflector support spars, and $|f(u, v)|$ clearly shows the uniform illumination resulting from the shaped reflectors and the spar blockage.

4.2. Pointing accuracy

The pointing accuracy of an antenna often limits the maximum usable frequency as much as reflector surface accuracy does. The desirable goal for the pointing accuracy, $\Delta\theta$, is $\Delta\theta < \theta_{3\mathrm{dB}}/20$ at the highest frequency of operation, where $\theta_{3\mathrm{dB}}$ is the full width to half maximum power of the antenna beam. With this performance, a source located at the center of the primary beam will suffer negligible intensity variations because $\mathcal{A}(\theta_{3\mathrm{dB}}/20) \approx 0.995$. As well as the on-axis intensity variations caused by pointing errors, the effect on a source located well out in the beam is important if extended sources are being imaged. With one-twentieth of a beamwidth pointing errors the fractional intensity variation of a source located at the half power point is $2\mathcal{A}(\theta_{3\mathrm{dB}}(\frac{1}{2} + \frac{1}{20})) \approx 0.87$, which will significantly reduce the accuracy of the outer part of the image.

The VLA antenna pointing accuracy is approximately 15 arcseconds. This corresponds to $\sim \theta_{3\mathrm{dB}}/7$ at $\lambda = 1.3$ cm, which is quite marginal since it will cause 5% intensity fluctuations at the beam center and 36% variations for a source at the half power point on the beam. One must consider the effect of this pointing error on wide-field images even at low frequencies. For example, at $\lambda = 21$ cm this pointing error causes 2% intensity variations at the half power point, which will be significant for high dynamic range images.

The design of the antenna for good pointing performance requires attention to a large number of details. Structural design for gravitational, wind and thermal loading on the antenna must be done carefully, with the latter two effects being the more important because they are not repeatable and therefore cannot be calibrated. A significant problem is provision of a sufficiently accurate model of the wind and thermal conditions at the array site. As well as deformations of the mount and reflector structure, the positional stability of the feed and subreflector can have a significant impact on pointing (Rudge et al. 1982, p. 169). The accuracy of any bearings and the precision and repeatability of the position transducers on the two axes must be considered. The servosystem and electromechanical drive designs play a major role in pointing accuracy, with wind performance being the most difficult problem. Finally, it is important to have a sufficiently accurate model in the control computer for both repeatable antenna pointing errors and atmospheric refraction. The coefficients of the antenna model are determined by measuring, on a periodic basis, the pointing errors on point sources distributed over the sky, and refraction correction requires data from an accurate and reliable weather station.

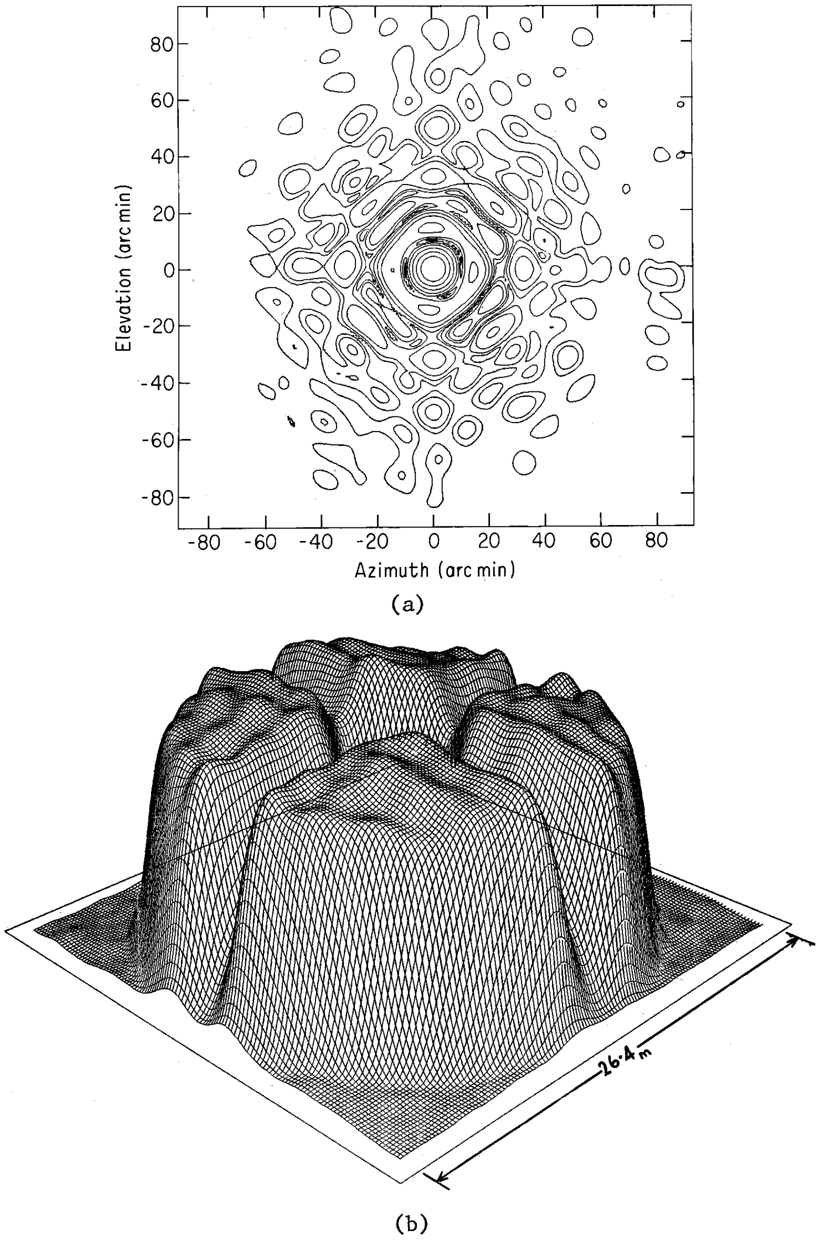
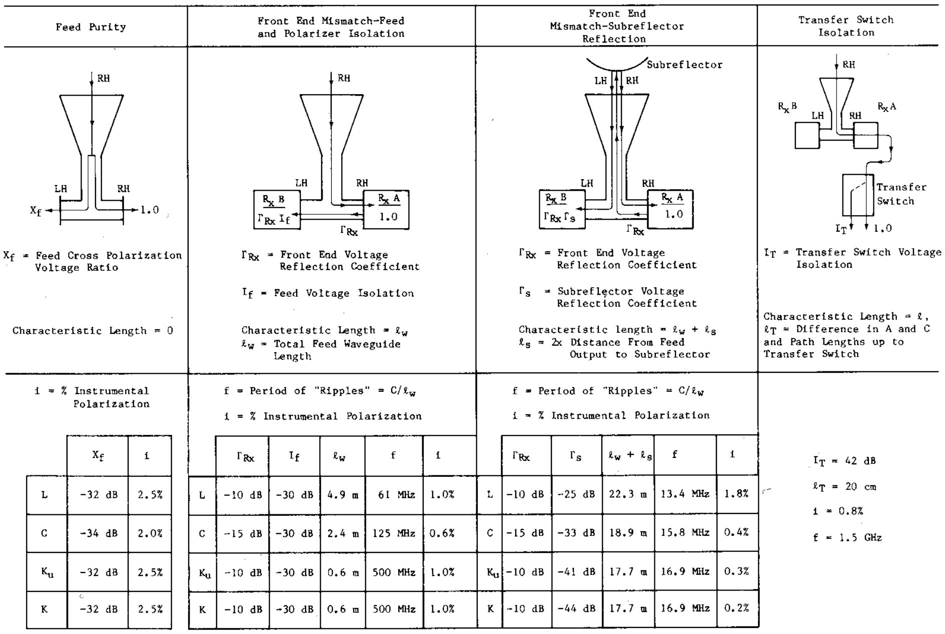


Figure 3–11. Results of a holographic measurement of a VLA antenna at 4.8 GHz. **(a)** $|F(l, m)|$ [from Napier, Thompson & Ekers (© 1983 IEEE)]; **(b)** $|f(u, v)|$. Phases are not shown because, at 4.8 GHz, they show little variation.



Feed Purity

X_f = Feed Cross Polarization Voltage Ratio

Characteristic Length = 0

i = % Instrumental Polarization

	X_f	i
L	-32 dB	2.5%
C	-34 dB	2.0%
K_u	-32 dB	2.5%
K	-32 dB	2.5%

Front End Mismatch-Feed and Polarizer Isolation

Γ_{Rx} = Front End Voltage Reflection Coefficient

I_f = Feed Voltage Isolation

Characteristic Length = ℓ_w
ℓ_w = Total Feed Waveguide Length

f = Period of "Ripples" = C/ℓ_w

i = % Instrumental Polarization

	Γ_{Rx}	I_f	ℓ_w	f	i
L	-10 dB	-30 dB	4.9 m	61 MHz	1.0%
C	-15 dB	-30 dB	2.4 m	125 MHz	0.6%
K_u	-10 dB	-30 dB	0.6 m	500 MHz	1.0%
K	-10 dB	-30 dB	0.6 m	500 MHz	1.0%

Front End Mismatch-Subreflector Reflection

Γ_{Rx} = Front End Voltage Reflection Coefficient

Γ_s = Subreflector Voltage Reflection Coefficient

Characteristic length = $\ell_w + \ell_s$
ℓ_s = 2x Distance From Feed Output to Subreflector

f = Period of "Ripples" = C/ℓ_w

i = % Instrumental Polarization

	Γ_{Rx}	Γ_s	$\ell_w + \ell_s$	f	i
L	-10 dB	-25 dB	22.3 m	13.4 MHz	1.8%
C	-15 dB	-33 dB	18.9 m	15.8 MHz	0.4%
K_u	-10 dB	-41 dB	17.7 m	16.9 MHz	0.3%
K	-10 dB	-44 dB	17.7 m	16.9 MHz	0.2%

Transfer Switch Isolation

I_T = Transfer Switch Voltage Isolation

Characteristic Length = ℓ,
ℓ_T = Difference in A and C and Path Lengths up to Transfer Switch

I_T = 42 dB

ℓ_T = 20 cm

i = 0.8%

f = 1.5 GHz

Figure 3–12. Sources of on-axis instrumental polarization using circularly polarized feeds for the VLA. [From Bignell 1982.]

Pointing performance can be improved if the pointing errors are frequently measured on a calibration source near to the object being imaged. In this observing mode, known as "offset pointing" or "referenced pointing" mode, corrections for the measured pointing errors are continuously applied in the computer pointing model for the antenna. Provided the pointing errors vary sufficiently slowly with time and antenna position, significant improvements in pointing accuracy can be made, generally to an accuracy of < 5 arcseconds for the VLA..

4.3. Antenna polarization properties

At the output of the antenna feed a polarization splitter provides separate output ports for two orthogonal polarizations, either linearly or circularly polarized. A number of mechanisms, discussed below, will prevent the output from a polarization splitter port from providing purely only one polarization component and none of the orthogonal one. The ratio of the undesired orthogonal component to the desired one, expressed as a voltage ratio, is known as the "instrumental polarization", "polarization crosstalk," or "polarization leakage". Instrumental polarization, if its effect is not removed from the data, can have a major effect on the quality of a polarization image. It will make an unpolarized source appear polarized and alter the apparent polarization distribution of a polarized source. The antenna instrumental polarization has two main components, a component

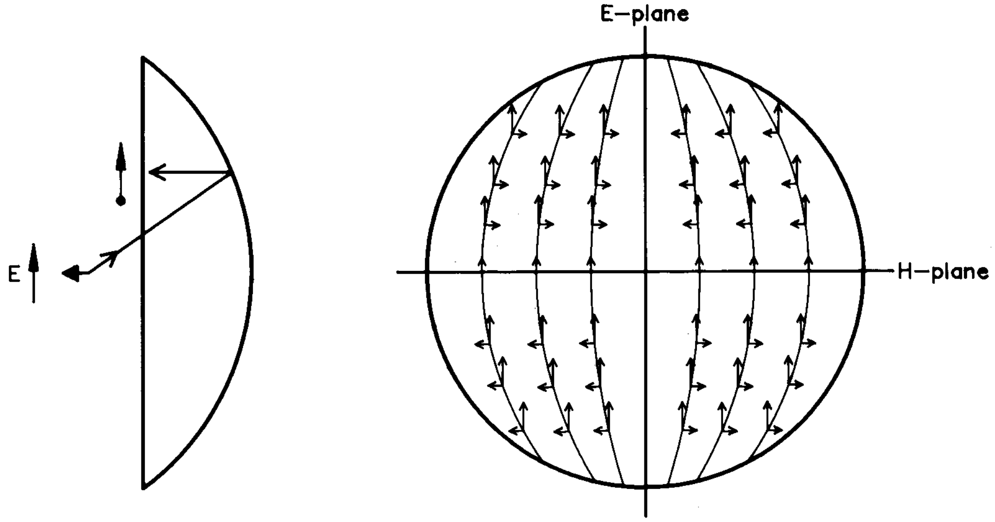


Figure 3–13. Field distribution in the aperture of a paraboloid fed by an electric dipole. The field is resolved into its co- and cross-polarized components.

at the center of the antenna beam that can be treated as constant across the beam and a component that varies across the beam. I will consider these two components separately.

The cross-polarization component that is constant across the beam is primarily due to the quality of the polarization splitter and, in the case of circular polarization, to reflections in the path between the subreflector and the receiver. The magnitudes of some of these effects for four of the VLA observing bands are shown in Figure 3–12. The goal with these effects is to make them as small as economically feasible and, more importantly, to ensure that they are constant in time so that their effect can be removed using the calibration procedure discussed in Lecture 5.

The cross-polarization component that varies across the antenna beam results from the curvature of the electric field lines in the feed pattern illuminating the main reflector (Rudge et al. 1982, p. 346). The situation is illustrated in Figure 3–13 which shows the electric field in the aperture of a parabola illuminated by an electric dipole. Note that the cross-polarized aperture distribution, $f_c(u,v)$ has the antisymmetric pattern shown schematically in Figure 3–14(a). The average value of $f_c(u,v)$ is zero, so it does not give rise to on-axis instrumental polarization. The cross-polarized radiation pattern, $F_c(l,m)$, which is the Fourier transform of $f_c(u,v)$, is also shown schematically in Figure 3–14(b). Note the four-lobed structure with the lobe peaks located at approximately the half power point of the primary copolarized beam. Typically these cross-polarized lobes are a few percent of the copolarized response at their peaks. An actual measured $F_c(l,m)$ for a VLA antenna is also shown in Figure 3–14. The effects of $F_c(l,m)$ are not removed by the usual polarization calibration techniques, and significant errors could be introduced into a polarized image of a source comparable in size to the antenna beam. The situation is further complicated on an alt–az antenna because $F_c(l,m)$ rotates with respect to the source,

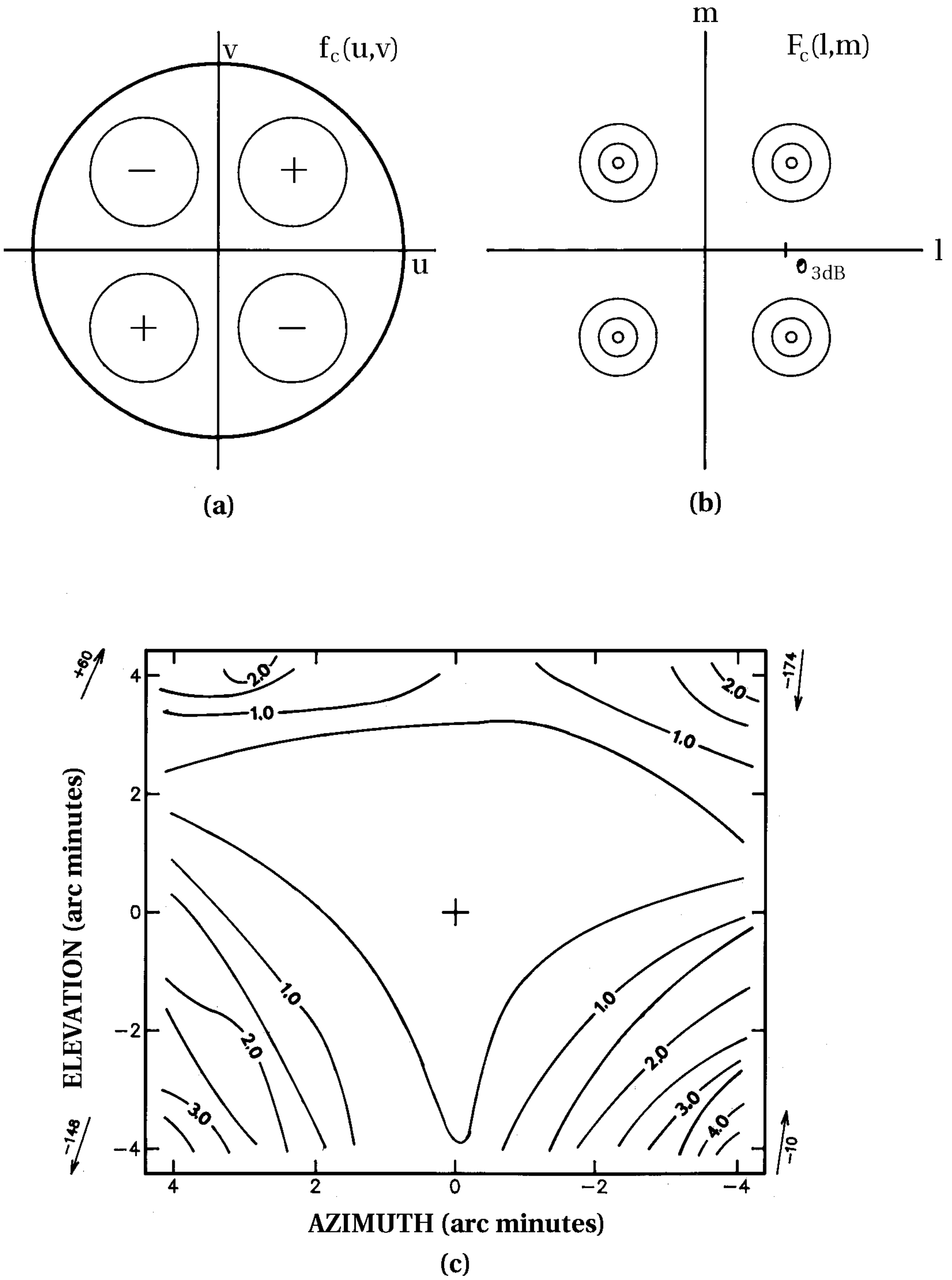


Figure 3–14. Cross-polarized beam of a reflector antenna. **(a)** Schematic diagram of $f_c(u,v)$ corresponding to the cross-polarized field in Figure 3–12. **(b)** Form of $F_c(l,m)$. **(c)** $F_c(l,m)$ measured on a VLA antenna at 4.8 GHz (from Bignell 1982). The half power point of the primary beam is 4.5 arcminutes. The on-axis instrumental term has been subtracted out, and contours are in percent polarization. The lack of symmetry of the four lobes results from the VLA asymmetric antenna geometry.

as explained in Section 3.2.. As well as the field curvature described above, four-lobed cross-polarized patterns of this type are also caused by non-circularity of the feed pattern and, to a small extent, by the curvature of the primary reflector. The peaks of $F_c(l,m)$ can be minimized by using multimode feedhorns, such as corrugated horns (Rudge et al. 1982, p. 359), which have circularly symmetric patterns and do not have the field curvature problems of a simple electric dipole.

A few words should be said about the effects of loss of symmetry in the antenna geometry. If there is no symmetry in the geometry then the anti-symmetry of Figure 3–14(a) will be lost and the cross polarization will not cancel in the on-axis direction. The off-axis geometry of the VLA, shown in Figure 3–6(c), has a plane of symmetry so the cancellation still occurs on-axis for linear polarization. Cross polarization in this case causes the circularly polarized beams to point in slightly different directions (Chu & Turrin 1973).

References

Balanis, C. A. 1997, *Antenna Theory*, Second Edition, (New York: John Wiley & Sons).

Bignell, R. C. 1982, *Polarimetry*, Lecture No. 6 in *Synthesis Mapping: Proceedings of the NRAO–VLA Workshop held in Socorro, New Mexico, June 21–25*, A. R. Thompson & L. R. D'Addario, Eds., NRAO (Green Bank, WV).

Chu, T. S. & Turrin, R. H. 1973, *IEEE Trans. Ant. Propagat.*, AP-21, 339–345.

Hansen, R. C., Ed. 1964, *Microwave Scanning Antennas*, Volume 1, Academic Press, New York.

Hudson, J., Plambeck, R., & Welch, W. J. 1987, *Radio Science*, 22, 1091–1101.

Kraus, J. D. 1986, *Radio Astronomy*, Second Edition, Cygnus-Quasar Books (Powell, Ohio).

Kraus, J. D. 1988, *Antennas, 2nd Ed.*, McGraw-Hill.

Mayer, C. E., Davis, J. H., Peters, W. L., & Vogel, W. J. 1983, *IEEE Trans. Instrum. Meas.*, IM-32, 102–109.

Napier, P. J., Thompson, A. R., & Ekers, R. D. 1983, *Proc. IEEE*, 71, 1295–1320.

Olver, A. D., Carricoats, P. J. B., Kishk, A. A. & Shafai, L. 1994, *Microwave Horns and Feeds*, IEE Electromagnetic Waves Series Vol 39.

Rudge, A. W., Milne, K., Olver, A. D., & Knight, P., Eds. 1982, *The Handbook of Antenna Design*, Volume 1, Peter Peregrinus, London.

Ruze, J. 1966, *Proc. IEEE*, 54, 633–640.

Scott, P. F. & Ryle, M. 1972, *MNRAS*, 178, 539–545.

Thompson, A. R., Moran, J. M., & Swenson, G. W., Jr. 1986, *Interferometry and Synthesis in Radio Astronomy*, John Wiley & Sons, New York.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

4. Cross Correlators

J. D. Romney
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. This lecture presents the general concepts of correlation and spatial coherence functions, as they apply to correlator implementations for synthesis-imaging arrays. A uniform treatment, based in the spectral domain, is applied to the both the conventional lag and the spectral or 'FX' correlator architectures. Parallel presentations of both architectures are given, including their strengths and weaknesses, although somewhat more emphasis is placed on the FX correlator because it is less familiar.

1. Introduction

The correlator of a synthesis-imaging array is the subsystem in which the interferometers are actually formed. Although a conceptual interference pattern may be considered to exist just from the act of receiving a common wavefront at the antennas of the array, it is in the correlator that interference fringes are formed and the complex visibility or spatial coherence function — introduced in the first lecture as the fundamental observable of synthesis imaging — is measured. Some have referred to the correlator as the "lens" of a synthesis array for this reason.

This lecture will address the concepts and theory of correlation in general. The theoretical treatment, based more in the spectral domain than has been customary, is designed to lead naturally to both the conventional lag correlator, and the spectral-domain or 'FX' architecture. Both connected-element synthesis arrays like the VLA, and VLBI arrays like the VLBA, are treated uniformly.

Following the spectral-domain basis of the theoretical treatment, and departing from previous treatments, I also will assume throughout that the desired output of the correlator is spectroscopic — i.e., a set of visibility measurements on a closely-spaced grid in radio-frequency space. This is clearly the requisite measurement for investigations of spectral emission or absorption lines; subsequent lectures will show that it is also the most appropriate form for constructing images of continuum sources, or for making high-precision astrometric or geodetic measurements.

2. Interfaces

The fundamental inputs to the correlator are digital samples from each antenna of the observing array. Lectures such as this sometimes cover details of the sampling process, but since the sampling occurs in other parts of the system, these details seem mainly to distract from consideration of the correlator itself, and in any case cannot be accommodated in the time available. Thus, I will take the position that the correlator's job is to measure the *sampled* signal.

The samples are obtained either in real time in connected-element arrays, or reproduced some time later from data tapes recorded in VLBI arrays. In the latter case, the correlator reproduces and resynchronizes the recorded samples, and thus recreates in one room the situation that existed across the world as the observations occurred. The delay between observation and correlation currently amounts to about two weeks for the VLBA correlator, although the range of turn-around times is limited in practice only by the performance of package-delivery services at the minimum, and the lifetime of a particular VLBI data recording system at the maximum.

The correlator's basic output is a set of measurements of the complex visibility function, spanning a total of four dimensions: the spatial-frequency vector (u, v, w) and the observing frequency ν. In the conventional ordering imposed by the correlator implementation, these results consist of spectra (over ν), for each member of the set of baselines, at a sequence of observing times. The baselines comprise all pairwise combinations of the observing stations, including "zero-baseline", single-dish measurements; observing time advances at intervals on the order of a second. Each baseline and time corresponds to a different (u, v, w) vector.

3. Background

At this point it's necessary to introduce some fundamental concepts and the formalism to be used throughout the lecture. The following subsections will have to summarize, in a very superficial fashion, some areas which are the subject of entire graduate-level courses, but I have given several references for further reading. And since I will be using mathematics principally as an heuristic tool, I will not pursue rigorous derivations.

3.1. Correlation and Coherence

The subjects of correlation and coherence can be approached from the point of view of optics (for which Born & Wolf 1980 is the definitive general reference), of communications engineering (where Blackman & Tukey 1958) is the classic work, although concerned only with what radio astronomers would call the "zero baseline" case), or of stochastic processes (see, *e.g.,* Parzen 1962). The terminology common in the latter field is closer to that of radio astronomy, and will be used in the following treatment.

We start with a quasi-monochromatic electromagnetic plane wave with an electric field component given by the real part of

$$E(t) = A(t)e^{+i2\pi\nu_0 t} \tag{4–1}$$

where $A(t)$ is a complex, band-limited, covariance-stationary, ergodic stochastic process. $A(t)$ can be considered to modulate a carrier at frequency ν_0. The wave propagates *opposite* to the unit vector pointing at the source, i.e., along $\hat{\mathbf{k}} = -\hat{\mathbf{s}}$.

Although $E(t)$ is an "analytic signal" as used in the usual complex-representation formalism, I will focus on $A(t)$, which is complex. Although the analytic-signal formalism thus will not be particularly useful, I will nevertheless work with complex functions because they permit a clearer presentation. Generally, physical signal processes will be the real parts of these functions, and I will point out the few cases where this somewhat careless usage requires correction.

We will be interested principally in the spectrum of $A(t)$, and so we introduce at this point its *spectral representation*

$$A(t) = \int_{-\infty}^{+\infty} s(\nu) e^{+i2\pi\nu t}\, d\nu. \tag{4–2}$$

It must be cautioned that $s(\nu)$ is strictly a mathematical device, and in fact has no physical meaning and is not measurable. Mathematicians are reluctant even to give it a name. Only when we reach Eqs. (4–6)–(4–7) below do we obtain what is usually thought of as a "spectrum". Nevertheless, we can still assert that $s(\nu)$ is zero outside a limited band $|\nu| \leq B$.

The interferometer measures what is called variously the cross-covariance, cross-correlation, or visibility function, defined as

$$\Gamma_{ij} = \langle E_i(t+\tau_{ij}) E_j^*(t) \rangle, \tag{4–3}$$

where $\tau_{ij} = (\mathbf{r}_i - \mathbf{r}_j) \cdot \hat{\mathbf{k}}/c$ is the light-travel time between observation of the same plane wave at the two spatial positions labeled i and j. The superscript * indicates complex conjugation.

In terms of the band-limited quasi-monochromatic signal $A(t)$,

$$\Gamma_{ij} = \langle A(t+\tau_{ij}) A^*(t) \rangle e^{+i2\pi\nu_0\tau_{ij}} = \gamma_{ij}(\tau) e^{+i2\pi\nu_0\tau_{ij}}, \tag{4–4}$$

where

$$\gamma_{ij}(\tau) = \langle A(t+\tau_{ij}) A^*(t) \rangle \tag{4–5}$$

is the covariance function of the stochastic process $A(t)$. The angle brackets in each of Eqs. (4–3)–(4–5) denote an expectation or statistical average. Since $A(t)$ has been declared to be ergodic, these can be approximated adequately by averages with respect to t.

Since τ_{ij} is realized by receiving the wave at different locations on the Earth, it will change as diurnal rotation changes the angle between the baseline vector and source unit vector; this gives rise to the "fringe" signal of interferometry which we must measure. Here $\gamma_{ij}(\tau)$ represents the complex amplitude and phase which modulate the fringe signal given by the phasor in τ_{ij}.

Using Eq. (4–2) we can develop a spectral representation for $\gamma_{ij}(\tau)$, as

$$\begin{aligned} \gamma_{ij}(\tau) &= \left\langle \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} s_i(\nu) s_j^*(\nu') e^{+i2\pi[(\nu-\nu')t+\nu\tau_{ij}]}\, d\nu\, d\nu' \right\rangle \\ &= \int_{-\infty}^{+\infty} S_{ij}(\nu) e^{+i2\pi\nu\tau_{ij}}\, d\nu, \end{aligned} \tag{4–6}$$

where

$$S_{ij}(\nu) = s_i(\nu) s_j^*(\nu) = \int_{-\infty}^{+\infty} \gamma_{ij}(\tau) e^{-i2\pi\nu\tau}\, d\tau \tag{4–7}$$

can now properly be called the "cross-power spectrum" of the process $A(t)$. A great deal of elegant mathematics has been devised to establish Eqs. (4–6)–(4–7) on a rigorous basis, even though the integrals in the Fourier transform relationship formally do not exist.

3.2. Observations

In this section I'll outline the major steps that occur at each observing station to produce the samples processed by the correlator. The quasi-monochromatic plane wave described by Equation (1) is received at station i, at position $\mathbf{R}_i(t)$, with a delay

$$\tau_i = \mathbf{R}_i(t+\tau_i)\cdot\hat{\mathbf{k}}/c \tag{4–8}$$

relative to the time the wave passes the coordinate origin. Note that τ_i must be defined recursively, in terms of the position of the station at an earlier time when the wave was received there. The coordinate origin relative to which $\mathbf{R}_i$ is measured is arbitrary, but is usually taken as some point near the center of a connected-element array, or the center of the Earth for a VLBI array. The time dependence of τ_i is not shown explicitly.

The received field

$$E_i(t) = E(t+\tau_i) = A(t+\tau_i)e^{+i2\pi\nu_0(t+\tau_i)} \tag{4–9}$$

is converted to baseband by a single-sideband mixer with LO frequency ν_0:

$$V_i(t) = E_i(t)e^{-i2\pi\nu_0 t} = A(t+\tau_i)e^{+i2\pi\nu_0\tau_i}. \tag{4–10}$$

This equation is not physically correct as written, since the LO signal is actually real, and the sum-of-frequencies component is removed by bandpass filtering. Nevertheless, the result is valid. A complete treatment also must account for frequency and phase offsets among the LOs at different stations, which I must gloss over here.

The baseband signal is sampled at an interval Δt, through a non-linear function $f[\cdot]$:

$$R_i(n) = f[V_i(t = n\Delta t)]. \tag{4–11}$$

Sampling will impose both a loss of sensitivity and a modification of the original spectrum. Techniques for correcting both effects in post-correlation analysis are well known. As I remarked at the outset, however, I consider here that the correlator's job is only to measure the sampled signal.

All subsequent manipulations of the signal operate on discrete, digital samples. However, most people find continuous functions and integrals more intuitive than series and summations, and I will proceed using the continuous forms whenever possible, which in fact will turn out to cover most of the material of this lecture. When necessary, I will resort to the discrete notation.

In a connected-element synthesis instrument, the digital samples can then be routed directly to the correlator. In a VLBI array, they are recorded on magnetic tape, in a format which includes embedded time tags to allow the time of each individual sample to be re-established. Offsets among the atomic time standards at the various stations (which can be kept reasonably small with modern technology) would have to be included in a more complete treatment.

4. The Correlator Frontend

At the input to a VLBI correlator system, the recorded samples are reproduced from the tapes, and the sample times re-established. It's sometimes possible to play back the tapes at a speed several times that used in recording, so that correlation can proceed faster than real time.

I have assigned to the correlator's "frontend" two other, essential functions, which compensate for the effects of the array geometry on the signals received. While it's not necessary (nor, is some cases, possible) to perform this compensation at the frontend of the correlator, it is convenient to introduce them conceptually at this point, because they follow naturally from the preceding section. At this point it's appropriate to discuss these functions in a completely station-based manner, although often they are thought of in a baseline sense. Subsequent sections will go into greater detail on their application in the two alternative correlator architectures, and will consider the baseline-based approach where necessary.

Implementation of these functions is an area where significant differences arise among specific correlator designs. Although a hierarchy of linear and/or polynomial approximations is almost always used, the number of coefficients and their update intervals vary widely.

4.1. Delay Compensation

Under the simplifying assumptions adopted earlier, compensation for the interferometer delay is extremely straightforward. Each reproduced signal $V_i(t)$ is delayed by δ_i:

$$P_i(t) = V_i(t - \delta_i) = A(t + \tau_i - \delta_i)e^{+i2\pi\nu_0\tau_i}. \qquad (4\text{–}12)$$

(Remember, the actual signal is a set of discrete samples of $V_i(t)$, $P_i(t)$, *etc.*) Ideally, one would choose

$$\delta_i = \tau_i \qquad (4\text{–}13)$$

to obtain the "white light fringe", as will be shown below. Departures from this ideal in different correlator architectures are discussed in subsequent sections.

The regular sampling implicit in Eq. (4–11) limits delay tracking at the correlator input to integral units of the sample interval. Techniques which can implement or approximate fractional-sample delay tracking will be discussed in subsequent sections. Connected-element interferometer systems, like the VLA, may implement a continuously-varying delay tracking by applying a timing offset to the sample clock, eliminating or minimizing the requirement for delay tracking in the correlator. This is feasible in a connected-element system where the correlator's operation can be embedded in that of the overall interferometer, but is substantially more difficult, although not impossible, in a VLBI correlator. Such a scheme was considered in the design of the VLBA, but was rejected as too difficult, and as an impediment to global VLBI compatibility.

In a VLBI instrument, it is possible to implement arbitrarily large delays simply by offsetting the tapes during playback. In practice, a digital buffer is necessary to accommodate minor variations in tape speed, so that at least a part of the delay can be, and inevitably is, tracked by manipulating the buffer pointers.

4.2. Phase Compensation

A phase rotation must also be applied to the signal, in order to stop the fringe signal as described earlier. We denote the phase compensation function as $\theta_i(t)$.

This phase rotation is almost always applied as a *complex* mixer, for reasons explained later. The *real* output of this rotator is then

$$X_i(t) = P_i(t)e^{-i\theta_i(t)} = A(t+\tau_i-\delta_i)e^{+i2\pi\nu_0\tau_i-i\theta_i(t)}. \qquad (4\text{–}14)$$

while the imaginary part is phase-shifted by $\pi/2$. Again the ideal phase rotation will be shown later to be correct:

$$\theta_i = 2\pi\nu_0\tau_i. \qquad (4\text{–}15)$$

Again, connected-element interferometer systems may eliminate the need for phase compensation in the correlator, by applying small offsets to the local oscillator frequencies in the baseband conversion step, similarly to the offset sampling described in the preceding subsection.

5. The Lag Correlator

With this background, we can consider the conventional or lag correlator architecture quite straightforwardly. Generalize Eq. (4–13) to provide a range of "lags", additional delay increments whose spacing, for the observing bandwidth B limiting the quasi-monochromatic process $A(t)$, is given by $\Delta\tau = 1/2B$:

$$\delta_i(n) = \tau_i + n\Delta\tau \qquad n \in [-\tfrac{N}{2}, \tfrac{N}{2}-1], \qquad (4\text{–}16)$$

and, using Eqs. (4–16) and (4–15) in Eq. (4–14), define a set of lagged signals

$$L_i(t,n) = A(t-n\Delta\tau). \qquad (4\text{–}17)$$

Then the correlation obtained by the lag correlator is

$$\begin{aligned} C_{ij}(n) &= \langle L_i(t,0)L_j^*(t,n)\rangle = \langle A(t)A^*(t-n\Delta\tau)\rangle \\ &= \langle A(t+n\Delta\tau)A^*(t)\rangle = \gamma_{ij}(n\Delta\tau). \end{aligned} \qquad (4\text{–}18)$$

Comparison with Eq. (4–5) shows that we have thus obtained the complex amplitude and phase which modulate the fringe signal in Eq. (4–4), and confirms that Eqs. (4–13) and (4–15) are correct. The N values of the cross-correlation function $C_{ij}(n)$ are now stationary, and can be integrated as long as desired, limited only by the range of residual fringe frequencies to be preserved.

Finally, to obtain the desired cross-power spectrum, it is then only necessary to apply the Fourier transform of Eq. (4–7). (This is a bit too glib; Eq. (4–7) implies $\gamma_{ij}(\tau)$ is sampled over an infinite range of τ, when in fact only N lags are obtained; the effect of this truncation is explored below.)

A M-station lag correlator, of which the VLA correlator is a large-scale example with $M = 27$, consists of $M(M-1)/2$ baseline units, each performing the complex multiplications and accumulations implied by Eq. (4–18). With only one primary unit type, arranged to form all possible station pairs, an overall block diagram is hardly necessary. A detailed view of the processing within each such unit appears in Fig. (4–1). The boxes labeled δ_i and θ_j represent the frontend delay and phase compensation, each applied to only one of the

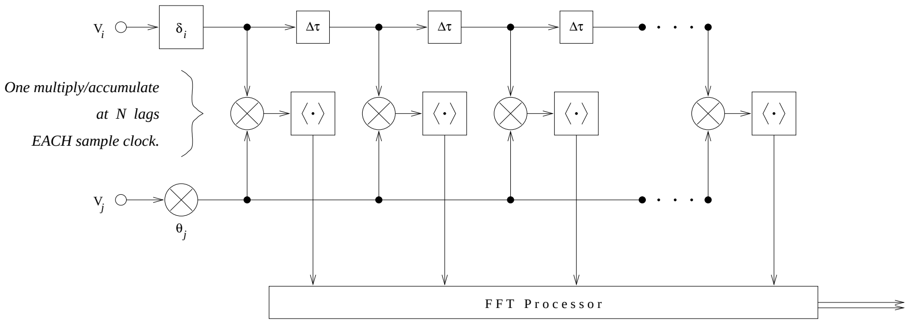


Figure 4–1. Lag correlator baseline processing.

station input streams v_i and v_j for reasons explained below. The individual one-lag delays marked $\delta\tau$ generate the lagged signals I have denoted as $L_j(t,n)$ in Eq. (4–17). These are multiplied by an unlagged signal corresponding to $L_i(t,0)$ in the mixer symbols, and the boxes denoted $\langle\cdot\rangle$ represent the statistical average to complete the implementation of Eq. (4–18). The final box is the Fourier transform corresponding to Eq. (4–7), although this is actually done off-line in many lag systems, including the VLA correlator. As indicated in the note at the left, each lag cell must be multiplied and accumulated on each sample clock, which will turn out to be a major contrast to the FX architecture.

5.1. Lag Correlator: Fractional-Sample Delay

I pointed out previously that delay compensation at the correlator input necessarily is restricted to integral samples. We are now in a position to see the effects of this, and to consider some solutions. Instead of Eq. (4–13), write

$$\delta_i = \tau_i + \varepsilon_i \tag{4–19}$$

where ε_i represents a fractional-sample error due to truncation or rounding of δ_i. Then in Eq. (4–17) we will have

$$L_i(t,n) = A(t - n\Delta\tau - \varepsilon_i), \tag{4–20}$$

and in Eq. (4–18),

$$C_{ij}(n) = \gamma_{ij}(n\Delta\tau - (\varepsilon_i - \varepsilon_j)). \tag{4–21}$$

On applying Eq. (4–7), the result of this lag shift will be to change $S(\nu)$ to

$$S'(\nu) = e^{-i2\pi\nu(\varepsilon_i - \varepsilon_j)} S(\nu). \tag{4–22}$$

Thus any fractional-sample error in delay compensation induces a *phase slope* across the observed band. Expressing the error in units of samples, f,

$$\varepsilon_i - \varepsilon_j = \frac{f}{2B}, \tag{4–23}$$

where B is the observing bandwidth,

$$S'(\nu) = e^{-i\pi f \frac{\nu}{B}} S(\nu). \tag{4–24}$$

To simplify the further analysis of this effect, we will assume, as is often the case, that the required delay compensation drifts through many sample intervals during the correlator's on-line integration time. Thus we regard f as uniformly distributed between some symmetric bounds, $f \in [-f_0, +f_0]$, and compute the expectation of the degraded correlator output spectrum:

$$S'_{f_0}(\nu) = \frac{1}{2f_0}\int_{-f_0}^{+f_0} S'(\nu)\,df = \frac{S(\nu)}{2f_0}\int_{-f_0}^{+f_0} e^{-i\pi f\frac{\nu}{B}}\,df = S(\nu)\cdot\frac{\sin\pi f_0\frac{\nu}{B}}{\pi f_0\frac{\nu}{B}} \quad (4\text{–}25)$$

The degradation of a continuum signal, $S(\nu) = S_0$, is obtained by integrating again, this time with respect to ν:

$$S'_0 = \frac{1}{B}\int_0^B S'_{f_0}(\nu)\,d\nu = \frac{S_0}{B}\int_0^B \frac{\sin\pi f_0\frac{\nu}{B}}{\pi f_0\frac{\nu}{B}}\,d\nu = S_0\cdot\frac{\mathrm{Si}(\pi f_0)}{\pi f_0} \quad (4\text{–}26)$$

where $\mathrm{Si}(\cdot)$ is the "sine integral" function.

We consider now some specific practical cases of f_0. Even if the error for each station is kept within $f \in [-\frac{1}{2}, +\frac{1}{2}]$ by careful rounding, independent station-based delay compensation can allow combined errors extending to $f \in [-1, +1]$ on a baseline basis. The effect of this is seen in Fig. (4–2a), where the two heavy lines show the phase shift, according to Eq. (4–24), across the observing band for the extreme values $f = \pm 1$, and the hatched area between them is the entire region covered for this case of $f_0 = 1$. From Eq. (4–25), we see that a spectral component at the upper band edge, $\nu = B$, is completely washed out, while the amplitude of one at the band center is reduced by $\sin(\frac{\pi}{2})/\frac{\pi}{2} = 0.637$. Across the entire band, Eq. (4–26) shows the amplitude of a continuum signal is reduced by $\mathrm{Si}(\pi)/\pi = 0.589$. These are all quite unsatisfactory.

Thus, lag correlators always require a fundamentally baseline-oriented delay compensation, where the rounding can ensure that $f \in [-\frac{1}{2}, +\frac{1}{2}]$. One common way to accomplish this is to use station-based delays with an extra, baseline-dependent *vernier delay* of $(0, \pm 1)$ bit. However implemented, this scheme achieves $f_0 = \frac{1}{2}$, as illustrated in Fig. (4–2b), with loss factors of 0.637 at the upper band-edge, 0.900 at the center, and 0.873 across a continuum band.

A further improvement requires coordination of the delay and phase compensation, inserting an extra phase shift of $\frac{\pi}{2}$ into the phase rotator whenever a one-sample delay shift occurs. Instead of Eq. (4–15) we have

$$\theta_i = 2\pi\nu_0\tau_i + \frac{\pi}{2}\frac{\delta_i}{1/2B} = 2\pi(\nu_0\tau_i + \frac{B}{2}\delta_i), \quad (4\text{–}27)$$

and Eqs. (4–20)–(4–22) become

$$L_i(t, n) = A(t - n\Delta\tau - \varepsilon_i)e^{+i\pi B\delta_i}, \quad (4\text{–}28)$$

$$C_{ij}(n) = \gamma_{ij}(n\Delta\tau - (\varepsilon_i - \varepsilon_j))e^{+i\pi B(\delta_i - \delta_j)}, \quad (4\text{–}29)$$

$$S''(\nu) = e^{-i2\pi[\nu(\varepsilon_i-\varepsilon_j)+\frac{B}{2}(\delta_i-\delta_j)]}S(\nu) = e^{-i2\pi[(\nu-\frac{B}{2})(\varepsilon_i-\varepsilon_j)+\frac{B}{2}(\tau_i-\tau_j)]}S(\nu). \quad (4\text{–}30)$$

In Eq. (4–27), since τ_i and δ_i are nearly equal, we see that the effect of the extra phase steps is to shift the phase tracking from the edge to the center of the

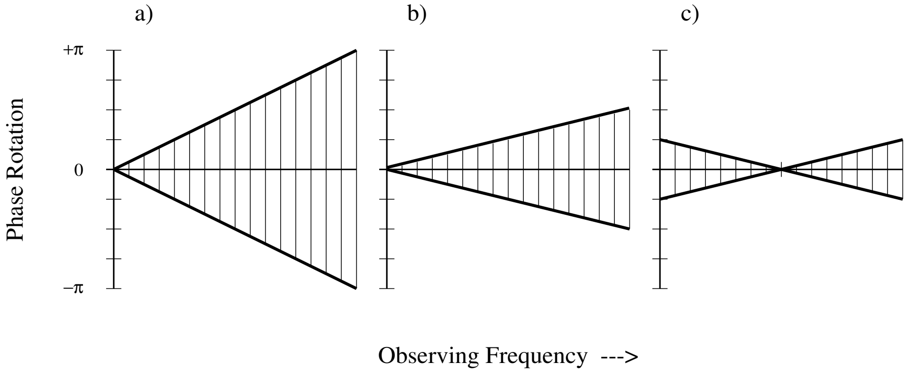


Figure 4–2. Phase shifts induced across the observing band for various configurations of delay and phase compensation.

observed band. This is reflected in the second term in the phasor of Eq. (4–30); in the first term, we obtain the desired change where the phase slope now pivots about $\nu = B/2$.

For this configuration, Eqs. (4–25) and (4–26) can be modified straightforwardly to

$$S''_{f_0}(\nu) = S(\nu) \cdot \frac{\sin \pi f_0(\frac{\nu}{B} - \frac{1}{2})}{\pi f_0(\frac{\nu}{B} - \frac{1}{2})} \tag{4–31}$$

and

$$S''_0 = S_0 \cdot \frac{\mathrm{Si}(\pi f_0/2)}{\pi f_0/2} \tag{4–32}$$

Assuming the baseline-based delay tracking introduced above, with $f_0 = \frac{1}{2}$, the phase shift variation across the band is as shown in Fig. (4–2c). Spectral components at the band center are unaffected, while those at *both* band edges are reduced by $\sin(\frac{\pi}{4})/\frac{\pi}{4} = 0.900$. The amplitude of a continuum signal is reduced by $\mathrm{Si}(\frac{\pi}{4})/\frac{\pi}{4} = 0.966$.

No matter which of these variants is used, it is necessary to apply some sort of correction in post-correlation software for the effects of integral-sample delay compensation in the lag correlator. Also, those cases where the assumption of rapidly changing delay fails can and must be corrected after correlation. These aspects of calibration are beyond the scope of this lecture.

5.2. Lag Correlator: Phase Rotator Implementation

A lag correlator must form a large number of lag products, since this must be done for every baseline, and so the efficiency of each lag is a critical design factor. This has limited lag correlator designs to at most four-level by four-level products until very recently. Thus with two- or four-level signal inputs, only primitive phase-rotation functions can be used. One very popular choice has been the three-level rotator (Clark, Weimer, & Weinreb 1972), shown in Fig. (4–3) along with the sinusoidal function which is being approximated.

Since the function is nearly a square wave, it is not surprising that a significant fraction of its power appears at frequencies other than the desired fun-

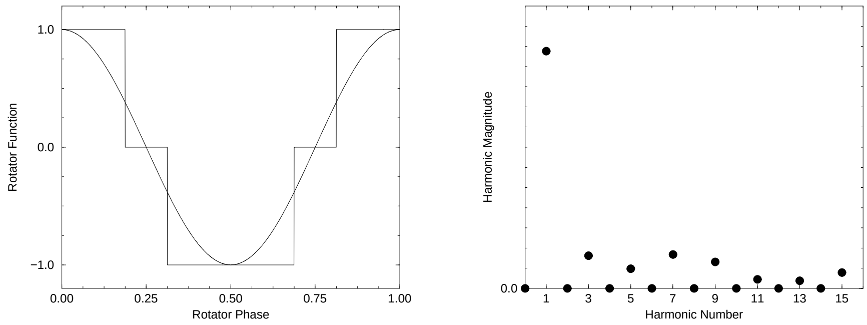


Figure 4–3. The three-level fringe rotator function, and its spectrum.

damental. Harmonic components will shift the input signals by the wrong frequency; these shifted components will not be stationary during an integration and this fraction of the input signal power will be lost. Consequently, lag correlators using this rotator function reduce the amplitude of the output fringe signal by a factor 0.96, which must be corrected in the calibration process.

Use of such a fringe rotation function has a further consequence for lag correlators. It is not possible to do the straightforward station-based fringe rotation implied by Eq. (4–15), for chance coincidences of different harmonics of two different fringe rotation functions will produce unexpected spurious correlations. Thus, as was true of the delay, phase rotation must be implemented on a baseline basis in a lag correlator.

5.3. Lag Correlator: Spectral Response

The limited range of lags, N in Eq. (4–16), has an important consequence in the profile of spectral lines observed with the synthesis instrument. The effect is fairly obvious, but should be developed here for comparison with the FX correlator architecture, where the corresponding situation is less well known.

We write $N\Delta\tau = T$, and represent the truncation of $\gamma_{ij}(\tau)$ as

$$\gamma'_{ij}(\tau) = \gamma_{ij}(\tau) \cdot \sqcap(\frac{\tau}{T}). \tag{4–33}$$

The "unit rectangle function" $\sqcap(x)$, an extremely useful device introduced in Bracewell's (1965) classic work on Fourier transform theory, is unity for $x \in [-\frac{1}{2}, \frac{1}{2}]$ and zero elsewhere. Using $\gamma'_{ij}(\tau)$ in Eq. (4–7), we can then apply the Convolution Theorem of the Fourier transform, and write immediately

$$S'(\nu) = S(\nu) * T \ \mathrm{sinc}(T\nu), \tag{4–34}$$

where the symbol $*$ indicates the convolution operation and the 'sinc' function $\mathrm{sinc}(s) = \sin(\pi s)/(\pi s)$, shown in Fig. (4–4), is the Fourier transform of $\sqcap(x)$.

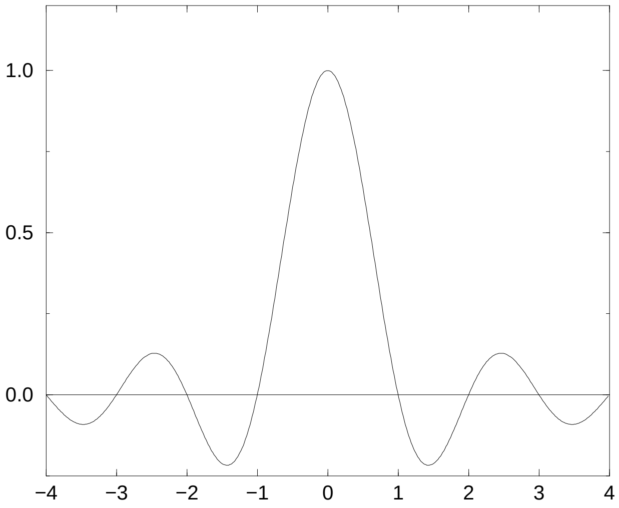


Figure 4–4. The sinc(·) function.

6. The FX Correlator

At the simple level adopted at the beginning of the preceding section for the lag correlator, the derivation of the spectral-domain or ‘FX’ correlator is almost too trivial to write down. If only as a framework for the description of the hardware stages, however, I will do so anyway. Starting again with the compensated input signal given by Eq. (4–14), for delay and phase compensation specified by Eqs. (4–13) and (4–15), we have directly

$$X_i(t) = A(t). \tag{4–35}$$

The FX correlator immediately Fourier transforms this signal — the ‘F’ of FX — to give

$$s_i(\nu) = \int_{-\infty}^{+\infty} X_i(t) e^{-i2\pi\nu t}\, dt \tag{4–36}$$

The FFT algorithm makes the entire FX concept practical, but is not a fundamental requirement. Next, these “station spectra” are *cross-multiplied* — the ‘X’ — yielding

$$S_{ij}(\nu) = s_i(\nu) s_j^*(\nu) \tag{4–37}$$

The cross-power spectrum in Eq. (4–37) can then be integrated. Again, the integration time is limited only by the desired range of residual fringe frequencies.

The name ‘FX’ was originated by Chikada et al. (1987), who also built the first such correlator, to indicate this reversal of the order of operations compared to the conventional lag correlator. The group at NRAO that designed the VLBA correlator considered one design of each type, and adopted the complementary term ‘XF’ which is sometimes used for the latter architecture.

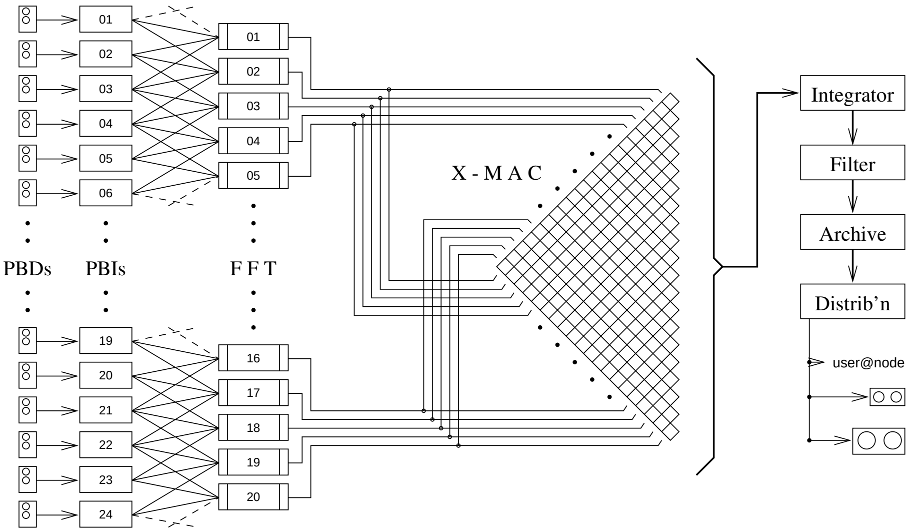


Figure 4–5. Overview of the VLBA Correlator, as representative of the FX architecture.

The VLBA correlator is presented here as representative of the FX architecture. Since this scheme is less familiar, and involves two basic unit types — rather than only one as in the lag correlator — an overview is shown in Fig. (4–5). The 24 VLBI tape playback drives (PBDs) and the 20 "playback interfaces" (PBIs) are only of interest here as implementing the integral-sample delay compensation. Phase compensation is performed in the unlabeled small boxes at the inputs to the 20 FFT processors, which themselves perform the 'F' operation of the correlator. The small boxes at the FFT outputs will be discussed in the next subsection. In the cross-multiplier/accumulator (X-MAC) a single square corresponds to the 'X' operation of the one-baseline correlator described here, and represents both Eq. (4–37) and a subsequent partial integration, which in turn is extended to longer times in the Integrator. Other components of the figure are not germane to the present discussion.

A detailed view of an FX correlator's baseline processing appears in Fig. (4–6). Here the frontend delay (δ) and phase (θ) compensation are shown applied separately to each station's input stream (v), since — as will be discussed below — this is feasible in the FX architecture. The "FFT processors" load segments of N points in a time series, each of which is Fourier transformed to produce "station spectra" which are then cross-multiplied, point by point, in the mixer symbol and accumulated. The accumulators yield the desired spectral-domain output directly after each integration cycle. As indicated in the note at the left, each spectral point must be cross-multiplied and accumulated only once for each N sample clocks — a factor of N lower rate than in the lag architecture.

Since FX correlators remain fairly unusual, I thought it would be useful to mention some, other than the VLBA correlator, that have been built. Following the original system built by Chikada et al. (1987) at Nobeyama Radio Observatory, Japanese scientists have built several other FX correlators, at Nobeyama and at the National Astronomical Observatory in Mitaka. Most notable among

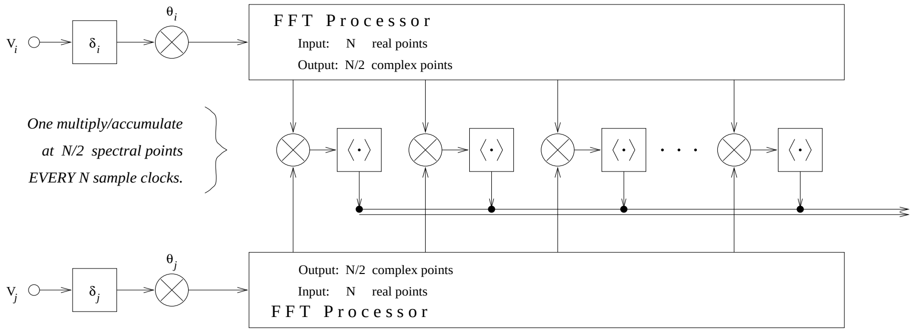


Figure 4–6. FX correlator baseline processing.

these, for the interests of local NRAO personnel, is the correlator built as part of the VSOP Space VLBI mission (Horiuchi et al. 1998). In the United States, a single-dish "Spectral Processor" was completed in 1990 at the NRAO Green Bank site. Some particular applications of that system are mentioned in subsequent sections where the relevant FX capability is discussed.

6.1. FX Correlator: Fractional-Sample Delay

The effect of integral-sample delay compensation at the correlator's front end is fundamentally the same for FX and lag correlators, but significantly different in the practical application. Again we use Eq. (4–19) to represent the integral-sample compensation with fractional-sample errors, and replace Eq. (4–35) with

$$X_i(t) = A(t - \varepsilon_i). \tag{4–38}$$

And similarly to the lag case, the FX correlator output is seen to be modified to

$$S'_{ij}(\nu) = e^{-i2\pi\nu(\varepsilon_i - \varepsilon_j)} S(\nu). \tag{4–39}$$

Although the same phase slope arises in both correlator architectures, in the FX scheme we can correct the slope at the Fourier transform output, before cross-multiplication and before integration. This operation is indicated in Fig. (4–5) by the small unlabeled boxes at the FFT outputs. Applying the correction at this point is equivalent to interpolating the fractional-sample delay at the correlator input. It allows the FX correlator to operate in all cases with *no fractional-sample loss at all.*

6.2. FX Correlator: Phase Rotator Implementation

Although the first stage of an FX correlator's Fourier transform could be optimized to accept the low precision (typically one or two bits) usually used in sampling, practical designs make all stages as similar as possible. Thus, the input typically can be presented at the same moderate precision as is needed at the Fourier transform's output. This allows the phase rotation to be done at that point using a multi-level rotator function, at essentially no extra cost.

Signals passed through such a rotator lose no significant power to harmonics. Thus, beyond just retaining the fringe power lost in a three-level rotator, it

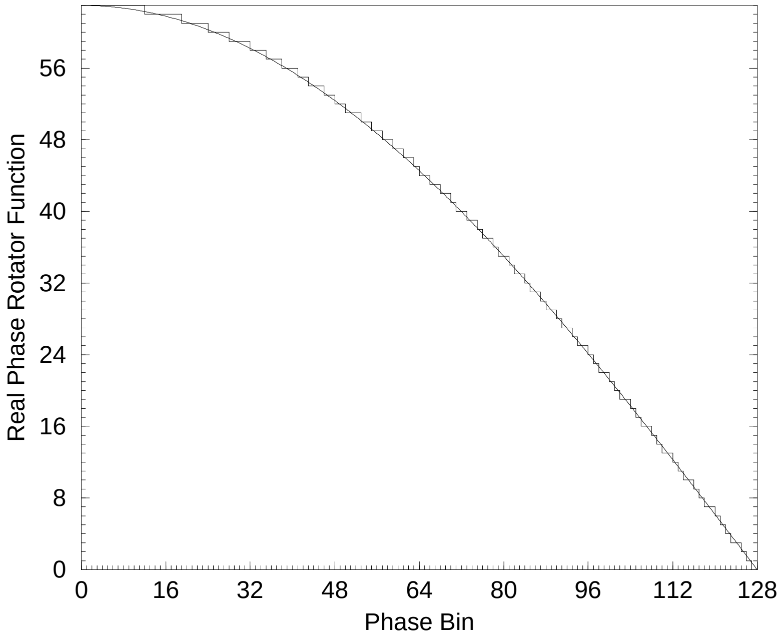


Figure 4–7. The VLBA correlator's multi-level phase rotator function.

becomes practical to implement station-based phase rotation without spurious correlations — and without the cost of a large number of baseline-based rotation elements.

Fig. (4–7) shows the multi-level rotator function used in the VLBA correlator — or actually just the real part of the first quadrant; all the rest is determined by symmetry. This function uses a 9-bit input phase word (spanning all four quadrants), and provides a 7-bit signed output value. This approximates the desired sinusoidal function so closely that only 5.3×10^{-5} of the input power goes into harmonics; no individual harmonic is more than about 25 dB below the fundamental.

6.3. FX Correlator: Spectral Response

Reversing the order of F and X operations has an interesting effect on the spectral profile. A practical FX correlator transforms a limited number, N, of samples in its F component — just as the lag correlator can form only a limited number of lags.

Proceeding as for the lag correlator, we write $N\Delta t = T$ (which is now an extent in time, rather than lag), and in place of Eq. (4–33),

$$X_i'(t) = X_i(t) \cdot \sqcap(\frac{t}{T}) \tag{4–40}$$

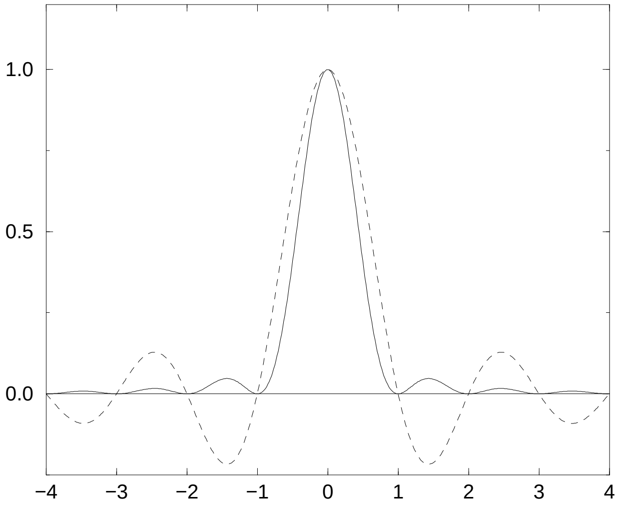


Figure 4–8. The $\text{sinc}^2(\cdot)$ function, with $\text{sinc}(\cdot)$ [dashed] for comparison.

The corresponding convolution in the spectral domain,

$$s_i'(\nu) = s_i(\nu) * T\ \text{sinc}(T\nu), \tag{4–41}$$

is applied to *each* of the station spectra, *before* cross-multiplication. Consequently, the cross-power spectrum is convolved twice by the $\text{sinc}(\cdot)$ function:

$$S_{ij}'(\nu) = S(\nu) * T^2\,\text{sinc}^2(T\nu). \tag{4–42}$$

The $\text{sinc}^2(\cdot)$ spectral response of the FX correlator, shown in Fig. (4–8), has both advantages and disadvantages compared to the $\text{sinc}(\cdot)$ response of the lag correlator (shown by a dashed curve in Fig. (4–8) for comparison). It is narrower, and has lower sidelobes, which generally are preferable characteristics for spectroscopy. Indeed, the FX-type Green Bank Spectral Processor has been exploited specifically to suppress interference signals. On the other hand, the narrow profile causes a greater "ripple" as a narrow line is scanned across the discrete spectrum. These characteristics are compared in the following table.

6.4. FX Correlator: Segmentation

In addition to the narrower spectral response, truncation of the input time series has a further effect on the performance of the FX correlator. This can be made evident by applying an inverse Fourier transform to Eq. (4–42) to derive the effective cross-correlation function measured by the correlator:

$$\gamma_{ij}'(\tau) = \gamma_{ij}(\tau) \cdot T \wedge (\tau/T). \tag{4–43}$$

The "unit triangle function" $\wedge(x) = \sqcap(x) * \sqcap(x)$ is the self-convolution of $\sqcap(x)$, having unit amplitude at $x = 0$ and tapering at unit positive and negative slope

Table 4–1. Spectral Response of Correlators

Correlator: Response:	Lag sinc(·)	FX sinc^2(·)
Full Width at 50% [spectral channels]	1.21	0.88
Full Width at 10% [spectral channels]	1.82	1.48
Height of First Sidelobe	−0.22	+0.05
Half-Channel Ripple	0.64	0.41

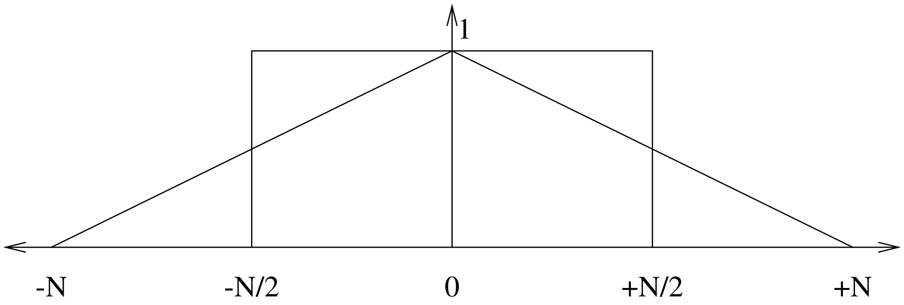


Figure 4–9. Relative density of lag measurements for lag and FX correlators.

to zero at $x = (-1, 1)$. Note that the width of $\wedge(x)$ is twice that of $\sqcap(x)$, while their integrals are equal.

Eq. (4–43) reveals two interesting characteristics of the FX correlator's response in the lag domain. First, transforming data samples in segments of length N yields a range of $2N$ lags covering $\tau \in (-N\Delta t, +N\Delta t)$. More importantly, however, these lags are heavily tapered by the triangle function; there are fewer pairs of samples within the set of N which can form large lags than small lags. Put another way, an FX correlator makes the same number of correlation measurements as a lag correlator with N lags, but they are distributed differently. The FX correlator reaches twice the lag range, but makes few measurements at the extremes of the range. This effect is shown schematically in Fig. (4–9).

The tapered distribution of independent cross-correlation measurements implies an oppositely-shaped distribution of noise in the measurements. The effect of this noise on a particular observation depends on the spectral characteristics of the source observed. A continuum spectrum has little power anyway beyond the first few lags away from zero, so the triangle tapering has little effect, while a narrow spectral line has significant power even at large lags. Before treating this effect of segmentation more quantitatively, we should consider an enhancement which can be applied when sufficient processing capacity is available: overlapping segments.

Operating with overlap factor f, an FX correlator advances its input data stream by only N/f samples for each N-point transform — requiring f times the processing capacity, of course. Some typical cases are shown in Fig. (4–10).

Overlapping does not change the triangle taper on the effective cross-correlation, nor the spectral response. It merely generates f times the number

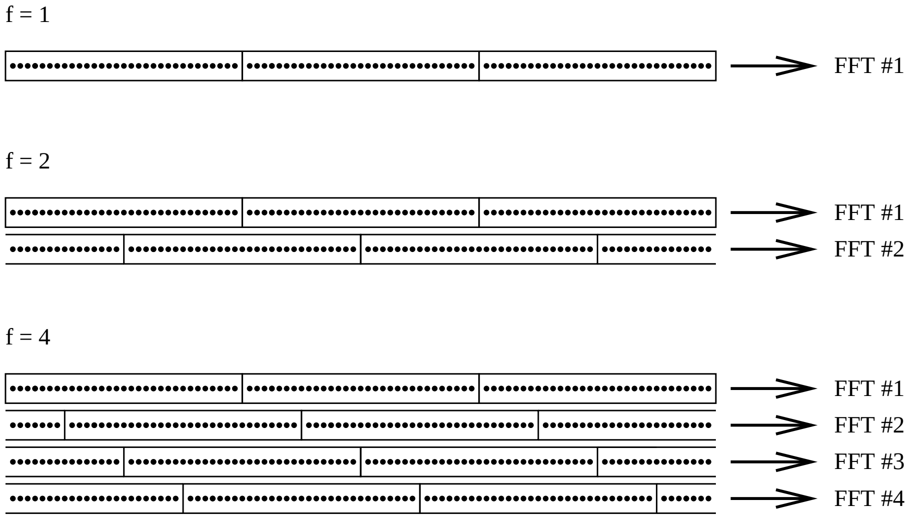


Figure 4–10. Cases of overlapped segmentation for $N = 32$ and $f = \{1, 2, 4\}$.

of correlation measurements. Some of these measurements are new: at large lags, $|\tau| \geq \frac{f-1}{f}N$, overlapping segments contain pairs of samples not otherwise combined. For smaller $|\tau|$, however, previously measured correlations are just repeated in the overlapping segments.

Thus we define a measurement weight

$$W(\tau) = \begin{cases} 1, & |\tau| \leq \frac{f-1}{f}T \\ f(1 - \frac{|\tau|}{T}), & \frac{f-1}{f}T \leq |\tau| \leq T \\ 0, & |\tau| \geq T \end{cases} \tag{4–44}$$

with $T = N\Delta t$ as previously. This function, a trapezoid with ever steeper ends as f increases (and degenerating to a triangle for $f = 1$), gives the fraction of all possible correlations which are measured at each lag τ. Fig. (4–11) shows $W(\tau)$ for some practical cases of f.

Then, for cross-correlation amplitude given by $C(\tau)$, we can adopt this relative measure of signal-to-noise ratio:

$$R = \frac{\int_{-T}^{+T} W(\tau)C(\tau)\,d\tau}{\int_{-T}^{+T} C(\tau)\,d\tau}. \tag{4–45}$$

R is normalized to unity for the non-overlapped case $f = 1$ which yields the same number of correlation measurements as a lag correlator.

Finally, we can evaluate Eq. (4–45) in two limiting cases. For a continuum source, $C(0) = C_0$ and vanishes elsewhere, and

$$R_{\text{continuum}} = 1, \qquad \text{for any } f. \tag{4–46}$$

And for a monochromatic line, $C(\tau) = C_0$ for all τ;

$$R_{\text{line}} = \int_{-T}^{+T} W(\tau)\,d\tau = \frac{2f-1}{f}. \tag{4–47}$$

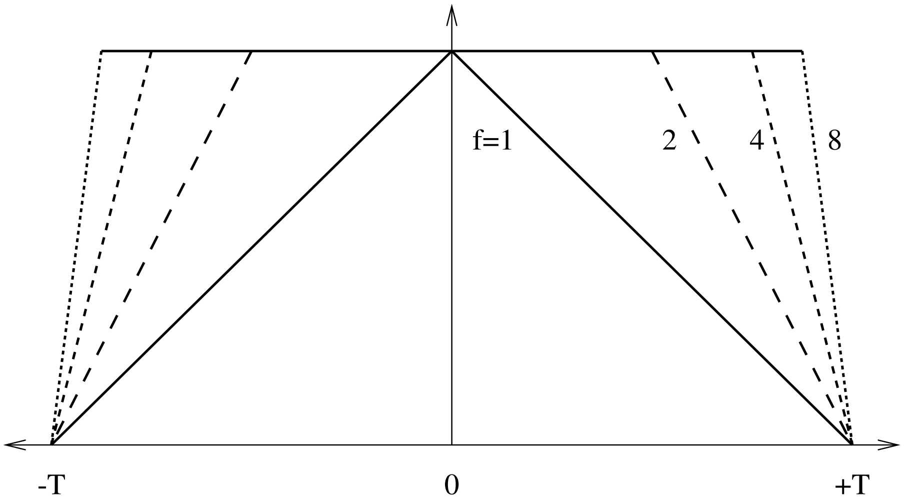


Figure 4–11. The measurement-weight function $W(\cdot)$ for some practical cases of f.

6.5. Miscellaneous FX Topics

Concluding my discussion of the FX correlator, I'd like to mention a few other aspects in which it differs significantly from the lag correlator. I'll start with a negative feature: it doesn't cope well with episodically invalid data. In a lag correlator, it's possible, at least in principle, to have a data-validity signal accompany each sample, and then inhibit correlation, independently at each lag, when either sample is invalid. In practice, this requires enough additional signal paths and normalization counters that it is seldom done — but in an FX correlator, it can't be done at all. Any practical FX design must use the FFT algorithm, which requires equi-spaced data samples. There is no way to omit an invalid sample. Generally this is not a problem at all for connected-element synthesis arrays. In VLBI arrays, the expedient of counting the errors for each FFT data segment, and invalidating the entire FFT output if the count exceeds some threshold, is usually satisfactory for typical recording and reproduction errors.

Next, I'll turn to two features which can only be implemented in an FX architecture. An input time-domain window can be applied at the same point as the phase rotation. Smoothing in the spectral domain, as commonly done in post-processing, typically applies only functions with a few points close to the peak, to minimize the computational load. With the input time-domain window, however, any real function can be used.

A pulsar gate can be applied at the FFT output, in the spectral domain. At that point, all delay and phase compensation has been performed, the signal had been resolved into narrow spectral bins, and an array-wide pulsar gate function — depending only on frequency and pulse phase — can be applied. The gate thus performs a form of de-dispersion impossible in a lag correlator, where the time-domain samples span an entire baseband channel. The Green Bank Spectral Processor uses this approach for single-dish pulsar de-dispersion. Such a gate was included in the design of the VLBA correlator, and in the original

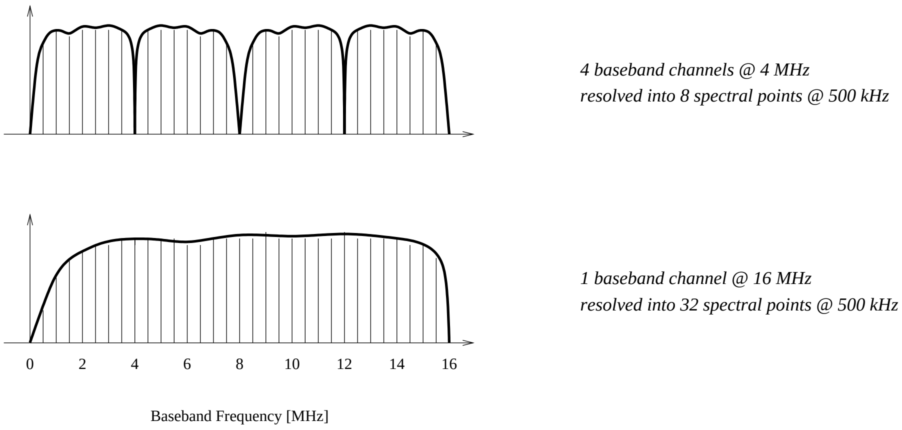

4 baseband channels @ 4 MHz
resolved into 8 spectral points @ 500 kHz
1 baseband channel @ 16 MHz
resolved into 32 spectral points @ 500 kHz
0
2
4
6
8
10
12
14
16
Baseband Frequency [MHz]

Figure 4–12. Example of an FX "hybrid mode".

hardware fabrication, but only became available recently with the completion of the necessary software.

Finally, we come to the dangerous "hybrid modes". It's possible to configure an FX correlator so as to transform, say, four contiguous baseband channels at relatively low spectral resolution, and cross-multiply the spectra against that from a single channel of four times the bandwidth, resolved into four times as many spectral bins. The individual bins from both sides thus match up, with the same frequency spacing. Such modes may be useful in VLBI arrays, to deal with incompatibilities among the disparate signal-processing and recording systems. Fig. (4–12) shows an example, which I must emphasize is purely conceptual, and has never been implemented. I've labeled these hybrid modes "dangerous" because, once people understand the concept, they expect any conceivable version to be achievable. In practice, each such case must be planned as part of the hardware design effort.

7. FX and Lag Correlator Intercomparison

To consolidate the material presented earlier, this final section will compare the two alternative correlator architectures from several quite different points of view. I'll begin with some very general considerations, then turn to a very particular basis for comparison — cost, and conclude by summarizing the other advantages and disadvantages of each.

Fig. (4–13) illustrates the four basic data types involved in interferometric correlation, and presents the lag and FX architectures graphically as two different routes from the common input, station sample streams, to the common result, baseline cross-power spectra. Two general conclusions should be apparent from this figure, as well as from the parallel derivations of the two correlators I presented earlier. First, the lag and FX correlators represent opposite sides of the Convolution Theorem of the Fourier transform: the transform of the convolution (correlation) equals the product of the transforms. And second, the FX architecture minimizes the number of operations required to perform the task,

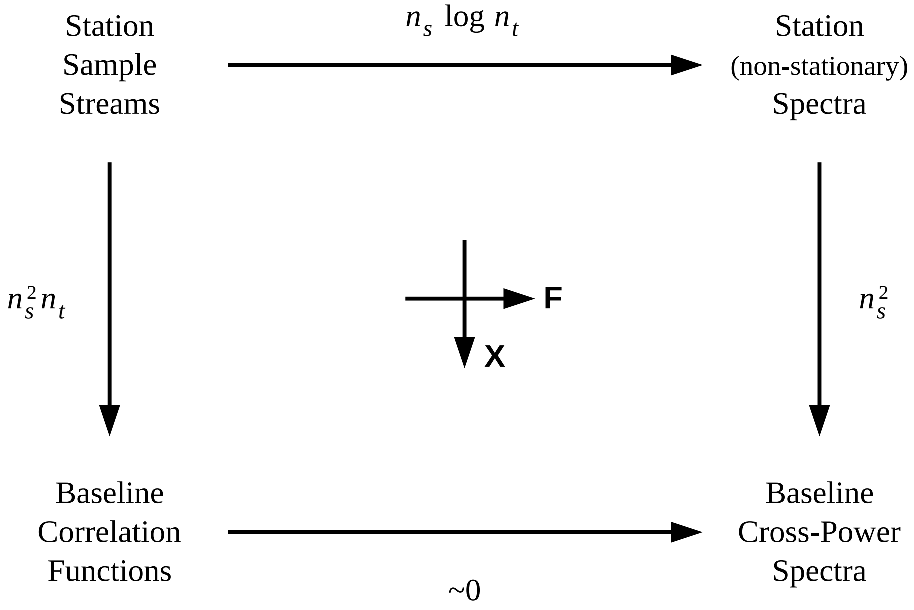


Figure 4–13. Comparative data processing paths for lag and FX architectures.

by exploiting the efficient FFT algorithm, and by organizing as much processing as possible on a station basis.

The figure shows, in terms of the system dimensions — n_s stations, and n_t samples transformed or lags formed — the operations needed along each of the four paths, and leads naturally into a more detailed discussion of the relative costs of the two alternatives.

7.1. Cost

Costs of the lag and FX architectures were compared during the design of the VLBA correlator. Two different approaches were followed, one based on simple operation counts (Romney 1986), and one on estimates of the required number of logic gates (Romney 1987). Both results are summarized here, for a correlator supporting n_s station inputs, each with n_c baseband channels (of sample rate r_0), and with n_t samples per segment transformed or lag correlations accumulated. The computations are based on $n_s^2/2$ baselines, including the real, single-dish "self-spectra".

The simplest approach to counting operations is just to count multiply operations. It can be shown that to a very good approximation, one addition accompanies each multiplication in both cases.

The lag correlator forms n_t lags per baseline and channel, each at rate r_0. As described earlier, phase rotation can only be performed on one input signal, so that each lag has one complex and one real input, requiring 2 multiplies. Thus the aggregate multiplier rate for cross-correlation is

$$r_L = r_0 n_s^2 n_c n_t. \qquad (4\text{–}48)$$

The contribution of the subsequent FFT operation can be neglected, since these lags can be accumulated for some time beforehand, so r_L is then the total multiply rate for the lag correlator system

For simplicity, assume the FX correlator's FFTs are implemented using a straightforward radix-2 Cooley-Tukey algorithm. Then the transform consists of $\frac{1}{2}n_t log_2 n_t$ "butterfly" stages, of 4 multiplies each, per station and channel. Transforms are executed at a rate r_0/n_t, so the aggregate multiplier rate in the FX correlator is

$$r_F = 2r_0 n_s n_c log_2 n_t. \tag{4–49}$$

The cross-multiplier must process $n_t/2$ points of each station/channel spectrum (because only one side of the spectrum contains signal power), at the same rate r_0/n_t. Since the spectra are complex, 4 multiplies are required per point. The aggregate cross-multiplication rate is then

$$r_X = r_0 n_s^2 n_c. \tag{4–50}$$

The ratio of multiplies for the two architectures,

$$R_{\frac{Lag}{FX}} = \frac{r_L}{r_F + r_X} = \frac{n_s n_t}{2log_2 n_t + n_s}, \tag{4–51}$$

shows that the FX architecture requires significantly fewer operations when *either* of n_s or n_t is large.

The gate-count approach was developed to circumvent one very serious oversimplification of the operation counts: the multiplies are of different complexities, with the lag correlator requiring only a few bits. Some serious preliminary work is required before a gate-count comparison can be attempted. Fairly thorough designs under both architectures, *using the same generation of microelectronic technology* (preferably as current as possible), must be developed.

From these one can extract the basic gate-count units: g_L gates per lag per baseline, g_F gates per FFT stage, and g_X gates per baseline cross-multiplied. It's important to ensure that the operations included in these basic gate-count are as functionally comparable as possible.

Then the lag correlator's total gate count (using the same system parameters as previously) is

$$G_L = n_c \frac{n_s^2}{2} n_t \cdot g_L, \tag{4–52}$$

and the corresponding total gate count for the FX correlator

$$G_{FX} = n_c n_s \lceil \log_4 n_t \rceil \cdot g_F + n_c \frac{n_s^2}{2} \cdot g_X. \tag{4–53}$$

Note that a radix-4-plus-2 FFT implementation is now assumed; the logarithm is to base 4, and the integer ceiling operation $\lceil \cdot \rceil$ accounts for the case where a radix-2 stage is required for n_t an odd power of 2.

Obviously, the results of such an analysis will depend sensitively on the gate-count factors g_L, g_F, and g_X, which can change drastically with time. Numerical results from the VLBA correlator study, done more than a decade ago, thus are no longer of any relevance. They did, however, show that the FX architecture would require fewer total gates, for any number of stations, for even quite modest spectral resolution — for the VLBA correlator's dimensions.

As developments in microelectronic technology make individual gates less expensive, however, other considerations may become more important, at least for very large-scale systems. The FFT stage of an FX system performs a data-*expansion* operation, and for the Millimeter Array currently under development it appears that the task of routing multi-bit signals to the cross-multiplier stage for an FX correlator would be prohibitive.

7.2. Summary of Other Advantages

The FX architecture allows a number of desirable features to be implemented which are not possible in a lag correlator. Some, such as truly station-based phase rotation and exact fractional-sample delay compensation, eliminate sensitivity losses as well as permitting more efficient implementation of essential functions. Others make unique, specialized observations possible (the spectral pulsar gate), or reduce the need for post-correlation processing (the time-domain window, and indeed the station-based phase rotation belongs in this category too). Its $\mathrm{sinc}^2(\cdot)$ spectral response may be superior, although it must be said there is as yet no experience with this. Although subject to segmentation loss, overlapped processing allows this to be restored, and to better sensitivity than an equivalent lag correlator.

The lag correlator is more tolerant of noisy transmission or recording of its input signals. And it actually obtains the cross-correlation function, which may be useful for certain cases of the correction for the non-linear sampling imposed on the observed signals.

References

Blackman, R. B., & Tukey, J. W. 1958, *The Measurement of Power Spectra* (New York: Dover).

Born, M., & Wolf, E. 1980, *Principles of Optics*, Sixth Ed. (Oxford: Pergamon).

Bracewell, R. 1965, *The Fourier Transform and Its Applications* (New York: McGraw-Hill).

Chikada, Y., et al. 1987, Proc. I. E. E. E., 75, 1203.

Clark, B. G., Weimer, R., & Weinreb, S. 1972, *The Mark II VLB System*, NRAO Electronics Division Internal Report 118.

Horiuchi, S., Kameno, S., Nan, R., Shibata, K., Inoue, M., Kobayashi, H., Murata, Y., Fomalont, E., and VSOG team 1998, *Advances in Space Research*, in press.

Parzen, E. 1962, *Stochastic Processes* (San Francisco: Holden-Day).

Romney, J. D. 1986, *Introduction to the Spectral-Domain ("FX") Correlator*, VLBA Correlator Memo No. 60.

Romney, J. D. 1987, *Spectral- vs. Lag-Domain Correlation: Comparison of Fundamental Hardware Requirements*, VLBA Correlator Memo No. 80.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

5. Calibration and Editing

Ed B. Fomalont
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A.

Richard A. Perley
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract.
The fundamental calibrations required to process radio interferometric data are discussed along with a description of the editing necessary to obtain a good calibration.

1. Introduction

Lecture 1 showed that, under certain well-justified assumptions, a two-element correlation interferometer measures the spatial coherence function of the radiation field at a location given by the antenna separation, or baseline, measured in wavelengths. Under other, generally well-justified assumptions, all these measurements of the coherence function can be considered to lie upon a plane, so that a two-dimensional Fourier transform suffices to obtain the sky brightness as a function of angle. (See Lecture 19 for a discussion of the consequences of having the measurements not lie sufficiently near a plane.) This spatial coherence function is usually referred to as the *true* visibility function and is denoted here by V_{ij}, with the subscripts indicating which pair of antennas is involved. An array of antennas samples this visibility function at many discrete locations. Lectures 2, 3 and 4 describe how the signals are collected, amplified, converted, transported, correlated and averaged. The data from each antenna pair are then recorded; the resulting ensemble of numbers is called the *observed* visibility data. In general, the recorded quantities are different from the desired visibilities. The intent of calibration is to recover the true visibility. We will use the symbol $\widetilde{V}_{ij}$ to denote observed visibilities.

The observed visibilities differ from the true visibilities for a multitude of reasons. The purpose this lecture is to discuss the origins and effects of the various mechanisms which affect the observed visibilities, to describe the methods used to measure the errors, and to describe how these derived parameters are used to estimate the true visibilities. We term this endeavor the *calibration* of the data.

The process of identifying and discarding discrepant and severely corrupted data is called *editing.* Discarding data is also commonly referred to as *flagging*, although the original meaning of the term was to note, or mark questionable data. Interference from terrestrial sources, antenna tracking inaccuracies, inclement weather, malfunctioning receivers, incorrect observing parameters, and data recording errors are but a few of the problems that may require flagging the affected data.

This lecture will emphasize calibration and editing problems and techniques in use at the VLA, with some digression on VLBI techniques. The discussion will emphasize general problems and not get too involved with specific VLA applications and software.

2. Basic Considerations

2.1. Synthesis equation and phase delay

The most useful array synthesis formulation for calibration purposes is Equation 1–13 from Lecture 1:

$$V(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}_\nu(l,m) I_\nu(l,m) e^{-2\pi i(ul+vm)}\, dl\, dm\,, \tag{5–1}$$

where

ν is the frequency of the radiation,
(l,m) specifies the direction cosines with respect to the phase-tracking center,[1]
(u,v) denotes the projected baseline coordinates, measured in wavelengths,
$V(u,v)$ is the visibility function, evaluated at (u,v),
$\mathcal{A}_\nu(l,m)$ is the normalized reception pattern of the antenna, and
$I_\nu(l,m)$ is the intensity distribution of the source.

Since the visibility is sampled at discrete times for each antenna pair, it is often useful to write the above equation as

$$V_{ij}(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}_\nu(l,m)\ I_\nu(l,m) e^{-2\pi i(u_{ij}(t)l+v_{ij}(t)m)}\, dl\, dm\,, \tag{5–2}$$

where we have explicitly expressed the visibility and the spatial frequency coordinates as functions of antenna pair (i,j) and time t. The term $(ul+vm)$ in the exponential is the geometric phase difference $\Delta\phi_g$ produced by the differential path length between the radiation from the source located at (l,m) to each antenna, compared with a fictitious source at the *phase-tracking center* with the assumed baseline between the two antennas.

The total geometric phase difference ϕ_g, or geometric group delay τ_g, for an interferometer with baseline components (L_x, L_y, L_z) is

$$\phi_g = 2\pi\nu\tau_g = \frac{2\pi}{\lambda}(L_x \cos H \cos\delta - L_y \sin H \cos\delta + L_z \sin\delta)\,. \tag{5–3}$$

[1] There are many 'centers' in radio interferometry, and we attempt to define them here. The *pointing center* is the direction of maximum antenna gain. The *phase-tracking center* is the direction for which the fringes have been stopped—that is, a point source in this direction will produce a constant measured phase (except for the influences of the atmosphere and electronics). This is also the direction for which the (u,v,w) coordinates are calculated, which thus defines the origin of the (l,m) coordinates. The *delay center* is that direction for which the coherence is maximized by inserting delay into one element of an interferometer to compensate for the geometrical and instrumental differential delay. It is important to note that these various centers are independent. The pointing and delay centers are fixed at the time of observation, but the phase-tracking center can be changed through recomputation of the baseline coordinates and phase-shifting the visibility data accordingly. This process is also known as moving the tangent point.

This equation follows directly from Equation 2–30 with the realization that the delay is simply $2\pi w$. Then, to first order, the differential geometric delay between radiation from another direction and the reference direction, or when the baselines differ from those assumed, is:

$$\begin{aligned}\Delta\phi_g = 2\pi\nu\Delta\tau_g &= \frac{2\pi}{\lambda}(\Delta L_x \cos H \cos\delta - \Delta L_y \sin H \cos\delta + \Delta L_z \sin\delta \\ &+ \Delta\alpha \cos\delta (L_x \sin H + L_y \cos H) \\ &+ \Delta\delta(-L_x \cos H \sin\delta + L_y \sin H \sin\delta + L_z \cos\delta))\,,\end{aligned}$$

where

$\Delta\tau_g$ is the differential geometric delay between the phase tracking center and the point of interest,
(L_x, L_y, L_z) is the assumed baseline separation for the antenna pair (i, j),
$(\Delta L_x, \Delta L_y, \Delta L_z)$ is the (true-minus-assumed) baseline,
(α, δ) is the true source position, and
$(\Delta\alpha, \Delta\delta)$ is the (true-minus-assumed) source position.

This relation is derived from Equation 2–31, using the astrometric coordinate grid shown in Figure 2–11. Note that an error in time is equivalent to an error in α.

Equation 5–3 is the basis for all interferometric analysis of antenna position coordinates and the basis for determination of astronomical positions.

2.2. Calibration formalism

An interferometric array generates the observed visibilities $\widetilde{V}_{ij}(t)$. In principle, the relationship between V and $\widetilde{V}$ can be arbitrary. Fortunately, however, adherence to sound engineering practices ensures that most arrays, to a good approximation, are linear devices: the output is a linear function of the input, or very nearly so. Furthermore, the individual measurements are well-isolated: the response associated with one antenna pair does not depend on the response of any other antenna pair. The basic calibration formula can therefore be written as

$$\widetilde{V}_{ij}(t) = \mathcal{G}_{ij}(t) V_{ij}(t) + \epsilon_{ij}(t) + \eta_{ij}(t)\,, \qquad (5\text{–}4)$$

where

t is the time of the observation,
$\mathcal{G}_{ij}(t)$ is the baseline-based complex gain,
$\epsilon_{ij}(t)$ is a baseline-based complex offset, and
$\eta_{ij}(t)$ is a stochastic complex noise.

Recall that the use of complex numbers is a convenience—it describes the combination of two correlator outputs (often termed the ‘Real’ and ‘Imaginary’, or ‘Cosine’ and ‘Sine’ correlator outputs) into one complex quantity. Thus the ‘complex offset’ and ‘complex noise’ are merely the complex resultants of the offsets and noises of two independent correlators.

In practice, the visibilities are averaged over a time interval during which they, or the complex gains, are not expected to change enough to perceptibly lower the coherence. Since the field-of-view of the antenna primary element limits the region from which signals are received, the maximum rate-of-change of phase will come from points at the edge of this region. It can easily be shown that an estimate of the maximum integration time in seconds is given by $\Delta t_{\mathrm{int}} \sim 2D_{(\mathrm{m})}/B_{(\mathrm{km})}$, where $B_{(\mathrm{km})}$ is the length of the baseline in km and $D_{(\mathrm{m})}$ is the diameter of the antenna in meters. Longer time intervals are permissible if the angular extent of the source is less than the primary beam size. The permissible integration-time scales with the ratio of the primary beam size to the source angular size.

2.3. Editing

Data editing can be considered a part of calibration, in the sense that there may be periods of time when $\mathcal{G}$ cannot be determined accurately for some or all of the antenna pairs. These data should then be removed.

There are several forms of editing: editing out the gross errors obvious in the uncalibrated data represents one extreme; and editing subtle, or suspected errors, where it is uncertain whether inclusion or exclusion of the data is the better choice, represents the other. Various handy software display tools aid in discovering the faulty data.

2.4. Calibration methods

Particular calibration methods depend on the detailed design of the array, the severity of the problem, and the ingenuity and desperation of the engineers and astronomers. Nevertheless, calibration methods can be divided into three basic categories.

(1) *Direct calibrations*: The tolerances in the design of modern arrays must meet exacting specifications to ensure stable and linear operation. For example, the VLA is designed to have amplitude stability of better than 1%, and phase stability of better than one degree of phase per gigahertz of observing frequency. Where instabilities and variations cannot be avoided, special feedback circuitry and the monitoring of critical parameters within the array system and in the environment around the array can be used to correct for these changes as the observations progress. Some examples are given in Section 4.

(2) *Calibrator sources in the sky:* An interferometer measures phase differences, so there is no absolute phase reference. For any given observation, we wish to reference the visibility phases to the phase-tracking center, which is generally the same position as the center of the primary antenna beam. To determine the antenna phase offsets, observations of a sky calibrator are required. Further, if the array is not completely phase- or gain-stable, periodic observations of calibrators are used to monitor these changes. Finally, the atmosphere will cause time-variable phase changes to occur in the data (mimicking the effect of unstable electronics), and observations of a calibrator source are often made in an attempt to remove this effect.

Calibrator observations are not concurrent or co-located in the sky with the actual observations.

(3) *Self-calibration:* In some cases the source being observed can be used as a test signal to calibrate the instrument. This type of calibration requires strong signals and particular array properties. This is the subject of Lecture 10 and will not be discussed in much detail in this lecture.

2.5. Calibration of amplitude and phase

A detailed look at the calibration Equation 5–4 and the use of calibrator sources will be given in Section 7. Most data corruption occurs before the signal pairs are correlated, so that the baseline-based complex gain $\mathcal{G}_{ij}(t)$ can be approximated by the product of the two associated antenna-based complex gains $g_i(t)$ and $g_j(t)$,

$$\mathcal{G}_{ij}(t) = g_i(t)g_j^*(t) = a_i(t)a_j(t)e^{i(\phi_i(t)-\phi_j(t))}\,, \tag{5–5}$$

where $a_i(t)$ is an antenna-based amplitude correction and $\phi_i(t)$ is the antenna-based phase correction. Observations of calibrator sources determine $\mathcal{G}_{ij}(t)$ for each of the $N(N-1)/2$ baselines, where N is the number of antennas. There are simple algorithms which then solve for the N values of $g_i(t)$.

2.6. Multi-frequency and dual-polarization capability

Many systems have multi-frequency (widely separated frequency bands) and dual-polarization capabilities. Most calibration operations that deal with the internal array system must be performed separately for each frequency and polarization channel, since the respective propagation paths differ. Calibrations associated with the geometry of the array or with the troposphere and ionosphere either are frequency independent or scale with known parameters. Such calibrations need to be done once, at a convenient frequency or polarization.

It is often useful to measure the visibility function at many frequencies within a narrow frequency range. These are called *spectral line* observations. These require an additional calibration, discussed in Section 5, in which the relative complex gains between the channels are determined. Except for this calibration, all of the other calibrations are identical.

If both orthogonal polarizations are obtained at one frequency, then there are four visibility functions that sample the spatial coherence function of the complete electromagnetic radiation field (see Sec. 5.2 of Lecture 1). Most of the correlation will be in only two of these—those including the I Stokes parameter. These correlations are used to calibrate the array parameters and the system gain. The remaining correlations (which depend on the linear and circular polarization of the field) require a special calibration, called *polarization calibration*, if polarization imaging is to be done. This calibration is discussed in Section 9.

3. Initial Calibrations

Before usable data can be taken, many instrumental parameters must be determined. These include the antenna pointing, delays and positions. Another related parameter is the accurate position of the calibrator to be used for monitoring system gain. This section describes these calibrations. In almost all cases, the parameters describing these effects must be determined and applied before useful observing can begin.

3.1. Antenna pointing and gain

The fundamental synthesis formulation of Equation 5–1 assumes that the primary beam distribution $\mathcal{A}_\nu(l, m)$ is independent of time and identical for each antenna. Small deviations occur, and some of the resulting effects can be corrected at later stages of observations. The antennas, of course, must be able to track accurately the diurnal motion of the source. Arrays composed of more than two elements with significantly different primary beam shapes can properly image only those sources which are small compared with the angular size of the primary beam, unless sophisticated imaging software is used.

For accurate imaging, the tracking of the center of the primary beam for all antennas must follow the intended sky position with an accuracy better than one-tenth of the full-width to half-power (FWHP) of the primary beam. Significantly larger errors can reduce the sensitivity of the observations and introduce distortions in extended objects. The 0.1 FWHP tolerance should be even smaller if a large part of the primary beam contains emission, either from one large-diameter source or from a high density of background sources.

The *antenna pointing error* is the difference between the actual pointing position (the location of the center of the primary beam) and the desired position. This difference has a complicated directional dependence because of misalignment of the polar or elevation axis, gravitational deformation of the structure, non-perpendicularity of the two axes, atmospheric refraction, and other effects (Clark 1973b). The antenna pointing error is also a function of time, because of differential heating of the structure. This pointing change can be estimated with tiltmeters placed on appropriate parts of the antenna structure, and it can be controlled to some degree by proper insulation of the antenna structure against solar heating. Wind-loading on the antenna also causes pointing errors, with time-scales of seconds.

Surveying methods and optical alignment of the antenna axis with Polaris and other bright stars can determine the pointing accuracy on the sky to within about ten arcminutes; better accuracy generally requires observations in many parts of the sky of radio sources of known position, to determine the directional dependence of the pointing errors. These are then fitted either to a physically-meaningful mathematical model describing the antenna defects, or simply to *ad hoc* trigonometric dependencies in the sky. If all of the pointing parameters are to be determined (typically five for each coordinate), then at least twenty observations, well-distributed around the sky, are needed to obtain sound estimates of the parameters.

Two types of observations are used to determine the pointing position from a calibration source, total-power observations and interferometric observations. Total-power observations have the advantage that they can be done by a single telescope observing bright extended sources. Interferometric observations have better stability, however. They also allow antenna-based pointing solutions, which provide greater accuracy.

The gain of an antenna decreases when observations are made near the horizon, because of deformation of the antenna structure and surface. Although it is often not a critical step in calibration, the dependence of gain upon zenith angle can be determined via the pointing error analysis. If the intensity of a calibrator source is known, then one can measure, as a function of zenith angle, the ratio of visibility amplitude to intensity. Often the antenna surface is designed so that maximum gain occurs at a zenith angle of about 30°. A decrease of more than a factor of two in the gain occurs only at high frequencies, near the limit of usefulness of the antenna.

3.2. Delay calibration

Equation 5–2 is the *monochromatic* synthesis equation. Since modern arrays are designed to operate over large instantaneous bandwidths, the frequency response over the desired bandwidth must be coherent. Large bandwidths, of course, are required to obtain sufficiently high signal-to-noise from the weak celestial sources. For a finite bandwidth, the integrated visibility function is

$$V_{ij}(t) = \int_0^\infty \left(\int_{-\infty}^\infty \int_{-\infty}^\infty \mathcal{A}_\nu(l,m) I_\nu(l,m) e^{-2\pi i \nu \Delta\tau_g}\, dl\, dm \right) e^{2\pi i \nu \Delta\tau_r} \mathcal{G}_{ij}(\nu)\, d\nu\,, \tag{5–6}$$

where

(i,j) denotes an antenna pair,
ν is the frequency,
$V_{ij}(t)$ is the visibility function integrated over the finite bandwidth,
$\mathcal{G}_{ij}(\nu)$ is the complex gain as a function of frequency,
$\Delta\tau_g$ is the differential geometric delay for the i–j baseline (delay relative to the delay-tracking center), and
$\Delta\tau_r$ is the residual instrumental delay for the i–j baseline (the error in inserted delay for the delay-tracking center).

If $\Delta\nu$ is used to denote the spanned bandwidth of the observations, then the phase difference $\Delta\phi$ between the ends of the band, that results from a net residual delay $\Delta\tau_g - \Delta\tau_r$, is given by

$$\Delta\phi = 2\pi\Delta\nu(\Delta\tau_g - \Delta\tau_r)\,. \tag{5–7}$$

A significant loss of coherence across the band will occur if $\Delta\phi$ is greater than about one radian. Thus, the delay between signals must be less than about $(160/\Delta\nu)$ nanoseconds, assuming $\Delta\nu$ is expressed in MHz.

Note that there are two different origins for a delay error. The first is geometric—only the delays for the delay-tracking center can be offset through insertion of delay in the system. Emission from other directions will suffer some loss of coherence. This the origin of the 'bandwidth-loss' effect noted in Lectures 2 and 18. The other origin is errors in the inserted delays, either due to mis-calculation or mis-calibration of the required delay, or due to the inevitable quantization of inserted delay.

Although the geometric delay can be precisely calculated, assuming knowledge of the antenna positions, a small residual delay error will generally remain, due to differences in signal propagation times through the different electronic paths. Note that these will vary between frequencies and polarizations within a single antenna, as well as between antennas. Delay calibration normally refers to the measurement of these small, but important, residual delays. The usual procedure is to observe a strong, isolated source of emission, and vary the delays in small steps until a maximum in the coherence is found. Another method, which in principle is much more sensitive, is to utilize the phase slope in frequency caused by a delay error (*cf.* Eq. 5–7). In this method, the delays are varied until the phase slope across the passband is zero. A spectral line correlator is required.

An often forgotten calibration is equalization of the delay between the two orthogonal polarization channels. Failure to correct for this delay will result in decorrelation of the cross-hand response, which normally provides the information on source polarization. Calibration of this delay is straightforward: observe a strong calibrator that is highly polarized (3C 286 or 3C 138 is commonly used), and adjust the delays until maximum signal is obtained on a cross-channel visibility function. This additional delay is the propagation difference between the two orthogonal polarization channels, and this amount should be added into all antennas of one polarization channel. It is best to do this after the parallel-hand delays have been set.

3.3. Time and place

The fundamental formula (Eq. 5–1) depends on the direction of the radiation (time, since the Earth rotates) and on the separation of the two antennas (place). If the baseline separation of the two antennas is in error or the source is not at the phase-tracking center, or if there is an error in the time, then the observed visibility phase $\varphi_{ij}(t)$ will vary with time, as given in Equation 5–3. The time reference, antenna locations, and calibrator location must be sufficiently well-determined that the visibility phase does not vary by more than about a radian over a characteristic time-scale. For the purpose of calibration, this time-scale will be the time between calibration observations, say, ten minutes. (This is a rather short time—one would prefer to lengthen the interval in order to increase integration time on the target source. Lengthening the time interval will make the baseline and time conditions more stringent.) A more fundamental time-scale is the basic integration interval, typically 10 to 30 seconds. Errors in these basic parameters which cause the phase to significantly change within this time window will cause serious error in the measured coherence. Another fundamental problem is in the calculation of the projected baseline (u, v, w). Incorrect antenna positions or time will cause errors in (u, v, w), and thence errors in the image due to incorrect gridding. The importance of this error will depend on the field-of-view and the required fidelity. See Section 3 of Lecture 19 for an estimate of the required accuracy. Finally, another ramification of a clock error will be erroneous pointing of the telescope. An error of a few seconds can cause the telescope to completely miss the source.

It is instructive to calculate the necessary accuracies in the baselines and in the time to keep the phase changes to less than a radian over 10 minutes. This condition requires the initial calibration of the baseline to be accurate to about four times the wavelength of the radiation, independent of the baseline length, or about 24 cm error at 5 GHz frequency. The accuracy of the time depends on the baseline length. For a frequency of 5 GHz, on a baseline of 10 km, a time error of 0.3 seconds will produce a phase error of about one radian over ten minutes. Smaller corrections are discussed in Section 7.7. Modern timekeeping can maintain the time to within a few milliseconds, equivalent to ~ 0.1 arcsecond. This is sufficient for coherent observations even with 30 km baselines at 24 GHz.

Calibrator sources with accurately known positions which are precessed accurately (see Clark 1973a, Clark 1982a) are observed over the sky in order to determine the antenna positions and timing error. For multi-element arrays, the data used are phase-referenced to a single antenna, whose position is assumed known. The errors $(\Delta L_x, \Delta L_y, \Delta L_z)$ in the baseline separations relative to this antenna, those in the source positions $(\Delta\alpha, \Delta\delta)$, and the time error Δt (which is equivalent to a zero-point shift in right ascension) can be determined by fitting the time and angle variations of these phases to Equation 5–3 (*cf.* Wade 1970; Brosche, Wade & Hjellming 1973). If the antenna locations are in error by many wavelengths, observations of closely spaced sources should be performed first. Alternatively, observations at a much lower frequency should suffice.

3.4. Amplitude check

A trivial but important initial calibration is to compare the approximate visibility amplitude from a radio calibrator for all of the baselines. If the calibrator is unresolved, the amplitude should be about equal for all baselines. If one antenna consistently shows a low amplitude with all its baselines, check the r.m.s. noise while observing a weak source. If the ratio of signal to noise is normal, then the antenna is probably working properly, but with the amplification somewhat different. If the signal-to-noise ratio is more than a factor of five lower than that for other antennas, then that antenna must be malfunctioning, and the data may not be useful.

4. Routine Corrections

The initial calibrations described above are generally applied on-line. Parameters describing the pointing characteristics of each antenna, the delay centering, the antenna locations, and the clock setting are inserted into the appropriate control computer. These allow the antennas to accurately track the source, and the phase and delay compensation for the phase-tracking center to be correctly calculated. However, these are not the only adjustments made while observing is in progress. Other corrections, which we call *routine corrections* are performed during observations. Although most observers do not deal with the routine corrections, these are important in determining the quality of the final image. These corrections include adjustments needed to track the phase and amplitude variation introduced by the electronics and signal transportation systems, and adjustments necessitated by the presence of the atmosphere. This section describes these adjustments.

4.1. Special array systems

Most arrays use sophisticated internal sensing circuits to measure and correct temporal changes in the instrument as the observations proceed. Some of the most important examples follow:

(1) Automatic Level Control: It is important that digital correlators receive a constant power level from each antenna. Most modern arrays use automatic gain control (AGC, or ALC) to maintain this condition. This is equivalent to making the receiver gain, when multiplied by the system noise temperature, invariant. However, the latter quantity may change for perfectly natural reasons, such as increased ground- or atmospheric pickup, or due to moving from one region of the sky to another. Under these conditions, the receiver gain will change, causing the resultant correlation coefficient to change. It is desirable to monitor the variations in receiver gain, and this is done by injecting a small amount of noise into the receiver and synchronously detecting the increase. These variations can then be used to correct the measured visibility. Gain changes outside of the receivers (e.g., antenna gain changes with elevation, and atmospheric attenuation) must be calibrated separately.

(2) Round-trip phase measuring schemes: Section 7.2 of Thompson *et al.* (1986) describes several schemes designed to measure the phase path length along the cables and waveguide that carry the local oscillator signal from the master oscillator to each antenna. It is important to monitor the variations of this path, since variations of the phase of the local oscillator at each antenna will cause loss of coherence in the observed visibility if the changes are large enough. The path length changes are corrected in the observed visibility while the data are taken.

(3) Tiltmeters on the axes of the telescopes can monitor changes in the pointing direction caused by wind and by thermal deformations. With sufficiently fast feedback loops, the pointing position of the antenna can, in principle, be corrected to within a few seconds of arc (Dewdney 1987).

The VLA and other arrays continually monitor temperatures, voltages, and other parameters in all parts of the system. When some parameters are outside of normal operating ranges, automatic editing of the affected data can occur. As experience with an array grows, clever uses of the monitoring data can be devised to better calibrate and edit the data during the observations.

4.2. Externally determined parameters

Some calibration parameters can be obtained from government monitoring programs or from other observatories whose instruments might be better suited to determining the relevant information. Weather-related parameters will be discussed in the next section. Earth-rotation and Earth axis orientation are monitored by concerted VLBI observations; up-to-date parameters are available from, e.g., Carter *et al.* (1985). Accurate positions of radio sources which are suitable for use as calibrators are also determined by VLBI observations. The longer-term motion of the Earth's axis, precession and nutation are encompassed in accurate precession algorithms that have been developed over the last several years. Smaller effects of Earth-tides, ocean loading, solar gravitational bending and motion of tectonic plates are now readily available.

The absolute gain of most arrays cannot be measured to better than 5% accuracy, but the relative gain stability is usually very much better than this. Thus, the relative flux densities of sources can be measured accurately, often to better than 0.5%. The absolute flux densities of the strongest sources are measured by single antennas with stable, well-determined gains. These values are adopted in most array work. An unfortunate problem is that most of the strongest sources whose absolute flux densities are known are also heavily resolved by modern interferometers. The commonly used flux density references are 3C 286, 3C 48, 3C 147, and 3C 295. Of these, only 3C 295 can safely be assumed to be non-variable.

4.3. Path length changes in the troposphere and ionosphere

Propagation of radio signals through the Earth's atmosphere causes a modification of the phase of the signal due to refraction occurring within the medium. Accurate interferometry requires correction of the visibilities for this perturbation. The effects can be ascribed to two causes: the global, large-scale structure; and the turbulent, small-scale structure. The turbulent effects are difficult to determine from direct measurements—calibrator observations are only marginally effective—except through self-calibration techniques. Most of the large-scale effects can, fortunately, be predicted and corrected in real-time.

It is important to remember that the phase of an external signal measured by an interferometer on level ground is unaffected by propagation through a plane–parallel atmosphere. This is because such an atmosphere affects equally all signals propagating through it, so that although the antenna pointing must be corrected for the effects of refraction, the interferometer cannot sense the presence of the atmosphere. However, a real interferometer is not on level ground on a flat Earth. Phase corrections are required, because of two effects: First, the antennas will be located at different heights. Second, the separated locations of the antennas mean that the objects observed will be seen at different elevations. Even though these differences are very slight, they are enough to require significant correction.

Refractive effects are usefully considered to have two origins, a dry component, about 6 km in thickness and a variable wet component, about 2 km in thickness. The decrease in the propagation speed of the radio signal is about one part in 3000, and independent of frequency up to about 300 GHz. This produces an additional time delay of the signal through the troposphere of about 8 ns (at the zenith), which is equivalent to an extra path length of 2.3 meters. From ground-based observations of the pressure P_{tot}, water-vapor partial pressure P_v, and temperature T, the excess propagation path can be roughly estimated. An approximate formula is

$$L = 0.228 P_{\mathrm{tot}} + 6.3 w \,, \tag{5–8}$$

where

$P_{\rm tot}$ is the total ground-level pressure in millibars,
w is the vertical column water-vapor content above the array (measured in centimeters), and
L is the total zenith excess path in centimeters.

A typical value of $P_{\rm tot}$ is 1010 millibars (at sea level). The water-vapor content is highly variable, ranging from as low as 0.1 cm to perhaps 5 cm (see Lecture 28). These zenith excess-paths can be converted into a phase path length between two elements of an array at the zenith angle of the observations. For zenith distances less than about 70°, a $\sec z$ law can safely be assumed. For antennas separated by more than 100 km, in which case the tropospheric conditions are essentially independent, a simple difference of the two path lengths is taken. For antennas that are closer, as in connected-element interferometry, a good approximation of the differential excess path length ΔL is

$$\Delta L = \left(\Delta h + \frac{c\tau_g L}{r_0} \sec z \right) \sec z \,, \tag{5–9}$$

where r_0 is the radius of the Earth, Δh is the height difference between the two antennas, and τ_g is the geometric delay between the elements. The second term is typically 10^{-5} of the baseline length, or a few centimeters for baselines of a few kilometers. It obviously becomes much larger for low elevations, especially when the antennas have a large differential delay—e.g., when they observe at the same azimuth as the baseline. These corrections are usually made as the observations progress. At zenith distances greater than $\sim 70°$, the curvature of the atmosphere cannot be neglected, and the dependence on zenith distance becomes more complicated.

Although the wet component of refraction is a small fraction of the dry component, the irregular distribution of water-vapor means that the highly variable phases which we see are dominated by this component. In the last few years, radiometers have been built to measure the integrated water-vapor emission in small regions of the sky, so as to allow estimation of the wet-term from this integrated emission. See the discussion in Thompson, Moran & Swenson (1986) for further details. There is some hope that such emission measurements may produce estimates of the wet-term excess path accurate to a few millimeters.

The ionosphere is a magneto-active plasma in the region 60–2000 km above the surface of the Earth. It affects the propagation of radio waves in two ways: the signal is refracted by the ionosphere, and its plane of polarization is rotated (this effect is discussed in Sec. 9.3 and in Lecture 29). The excess path L_i (in meters) in the zenith direction is

$$L_i = -40\nu^{-2} N_e \,, \tag{5–10}$$

where ν is the frequency in GHz and N_e is the electron column density in units of 10^{18} m^{-2}. The excess path length is negative because the phase is *advanced* with respect to a wave *in vacuo*. At 5 GHz using a typical daytime value for N_e of 3×10^{17} m^{-2}, a path length difference of about 0.5 meter is obtained. A typical nighttime value of the electron density is about a factor of five less. The differential excess path ΔL_i between two antennas observing at zenith angle z is

$$\Delta L_i = \frac{L_i c \tau_g}{r_0 \cos^2 z + 2h}\,, \tag{5–11}$$

where h is the height of the ionosphere. An estimate of the electron column density in the ionosphere can be obtained from the average solar activity or can be measured using either an ionosonde or dual-frequency satellite transmissions.

The path length difference of the ionosphere is about the same as that of the troposphere at a frequency of about 1.4 GHz. Since the ionospheric refraction is dispersive (i.e., frequency dependent) it can be removed using dual-frequency observations.

4.4. Absorption by the troposphere and ionosphere

Atmospheric opacity has two effects on visibility amplitude: First, the strength S of the source signal is decreased to $Se^{-\tau_a \sec z}$. Secondly, the atmosphere's emission adds to the system noise a contribution equal to $T_{\rm atm}(1 - e^{-\tau_a \sec z})$, where τ_a is the zenith opacity of the atmosphere and $T_{\rm atm}$ is the average tropospheric temperature. Opacity can be determined by observing calibrators of known flux density at various elevations. However, the flux densities must be accurately known, the system gain must be very stable, and the dependence of antenna gain on zenith angle may cause confusion. Tracking a strong calibrator over a large range of zenith distance removes the first objection, but the second and third requirements remain. A better method is to use *tipping curves*, in which the total power from an antenna is monitored as the antenna slews from the zenith to the horizon. The dependence of total power on zenith angle provides a measure of opacity. Stable receivers are needed for accurate tipping curves (*cf.* Uson 1986).

An approximate estimate of the opacity can be obtained from ground meteorological measurements via the formula

$$\tau_a = \alpha_0 + \alpha_1 P_v\,, \tag{5–12}$$

where P_v is the water-vapor partial pressure in millibars. The following table lists α_0 and α_1 as functions of frequency.

Table 5–1. Coefficients to Estimate Atmospheric Opacity

Frequency	α_0	α_1
15 GHz	0.013	0.0002
22 GHz	0.026	0.015
35 GHz	0.039	0.0039
90 GHz	0.039	0.013

Ionospheric absorption of radio signals is minimal. At a frequency of 100 MHz during periods of high ionospheric activity, the absorption is about 2% and varies as the inverse-square of the frequency.

5. Bandpass Calibration

We have, so far, ignored finite bandwidths, except for the discussion of delay calibration in Section 3.2. There it was assumed that the observer was interested only in the integrated visibility over as wide a bandwidth as possible, in order to obtain the best signal-to-noise ratio. These are called *continuum* observations. It is acceptable to sum the emission over the bandwidth prior to correlation, provided that the delays are properly set (meaning the phase slope is very small), that there are no important changes in the visibility within the integrated band, and that the effects of chromatic aberration are unimportant.

However, these conditions may not be met, in which case it is necessary to generate the coherence as a function of frequency. In particular, many interesting radio sources contain atoms and molecules that emit most of their radiation in narrow frequency ranges, so it is important to image the source at a large number of adjacent frequency channels that span the width of the molecular emission, and with sufficient resolution in frequency to separate emission regions at different radial velocities. Or, in many cases, the effects of chromatic aberration cannot be ignored, due to the presence of strong background sources far from the phase-tracking center. In this case, observations must be taken with narrow bandwidths to allow a larger field-of-view. If the sensitivity is to be maintained, many observing bandwidths (or channels) are required. Such observations are called *spectral line* observations.

Delay calibration (Sec. 3.2) equalizes the propagation difference of the individual signal paths at the input of the correlator and removes the large phase gradient across the observing band that would otherwise occur. To handle changes of antenna gain with frequency, we can consider the baseline-based complex gain to be a function of frequency—$\mathcal{G}_{ij}(\nu)$. Compensating for the change of gain with frequency is called *bandpass calibration*. For a well-engineered system, variations of amplitude and phase across the bandpass will be small. Nevertheless, for many observations, the required channel-to-channel accuracy will be greater than can be provided without bandpass calibration. This section briefly describes these calibrations.

The relative frequency response of the set of frequency channels can be determined by observing a strong calibrator source for sufficient time to reach the required accuracy. It is important that the spectrum of the calibrator be flat over the frequency band. It need not be a point source. Since the true visibility is identical in all channels for a calibrator, the bandpass complex correlator-based gain function for the i–j baseline and the k^{th} frequency channel is the observed visibility $\widetilde{V}_{ij}(\nu_k)$, divided by the correct visibility $V_{ij}(\nu_k)$. The latter can be determined by observing a very strong source with a continuum correlator. The frequency-dependent, baseline-dependent gain can be factored into products of antenna-based gains, so that the set of baseline-dependent bandpasses can be converted into antenna-based bandpass amplitude and phase calibrations. The bandpass function is not a strong function of time and is determined relatively infrequently. In principle each significant change of frequency requires a new calibration, but the tolerance depends on the frequency characteristics of the array and the accuracy of the observations. Certainly the small frequency changes that are made to compensate for the rotation of the Earth are insignificant as far as the calibration is concerned.

Because of the narrow channel-width, accurate calibration of the antenna bandpasses should only be done using very strong calibrators. Fortunately, the time variation of these bandpass functions is very slow, so that for most applications, one calibration per observing run is sufficient. Calibration of other array parameters, such as temporal phase and gain fluctuations which affect all the frequency channels equally, is accomplished using the entire bandpass signal (i.e., summing the response over frequency). Again, because of the slow variation of the system parameters with frequency, this calibration can be done before the bandpass corrections are made.

In modern synthesis telescopes, the frequency separation is provided by a digital cross-correlation spectrometer, rather than by a set of narrow-band filters. The method and theory are discussed in Lectures 1 and 4. The output from the Fourier transform of the cross-correlated spectrum is a set of visibility functions at a grid of frequencies. The bandpass function can be obtained from the same calibrator observations described above. A minor problem introduced by the finite number of time lags in the correlator is that the ripples in the bandpass function depend on the input phase of the observed visibility! The real part of the visibility function is convolved somewhat differently than the imaginary part. This can be remedied by lowering the frequency resolution by smoothing (thus decreasing the amplitude of the ripples), or by increasing the number of frequency channels (which increases the number of measured time lags).

6. Editing

When parts of an array malfunction, data of poor quality are recorded, and their inclusion in the imaging process will deteriorate the results. Even if it is possible to calibrate the severely affected data, it may have too much noise or be unstable. Data of poor quality are usually worse than no data at all. It is impossible to display all of the data from modern synthesis telescopes. By carefully choosing a subset of representative data, it is usually possible to discover faulty data in nearly all of the data set, except in pathological cases.

Editing of array data is usually done in four steps. The first step was mentioned in Section 3.5—determining which antennas were working. This section describes the second step of editing: removing outliers and looking for inconsistent data. Part of this editing is done by the automated, on-line monitoring system and part by the observer looking over the data. The third step is done during the calibration process, when longer-term problems come to light and more subtle data problems may surface. The last editing step comes after the image is made. If this image is of poor quality (noisier than it should be or containing peculiar artifacts; see Lecture 13), then more data editing may be required or the calibration may be in error.

6.1. External monitoring

Modern synthesis telescopes have internal checks which are intended to determine if the data being collected are of good quality. If it is clear that the data are in error (for example, if an antenna is not pointed at the source), then the system should not even record the data. If there is any question about data validity, then the data should be recorded and flagged bad, rather than discarded.

6.2. Scan consistency

Most observations consist of scans where the visibility function is sampled every few seconds for a duration of several minutes. Since the visibility function for any source is a slowly varying function of time, relatively simple tests of the data consistency in a scan can be made. If the source visibility is limited by noise (each sampled data point $< 5\sigma$) then only bursts of interference can be detected in this manner. If each data point has good signal-to-noise ratio ($> 10\sigma$), then instabilities in the amplitude and phase can also be detected. The easiest test is to determine the r.m.s. spread of the data about the mean visibility (remember, it is a complex number) of the scan. Unless the source is strong, in which case slight gain instabilities will be the more important contribution to the r.m.s. scatter, it should be consistent with the expected noise. For a strong source, if the phase scatter from the troposphere is large, then it is better to look at the r.m.s. scatter of the visibility amplitude (rather than that of the complex visibility) to find discrepant points. But also keep in mind that any visibility variations of a strong source can easily be corrected through self-calibration. One wants to retain those data that can be corrected, and delete only those that are incorrigible.

The second method can be used if both polarizations at one frequency are simultaneously observed with a telescope; then the difference between the two parallel-hand visibility measurements should be consistent with noise. All atmospheric fluctuations should cancel. The difference in the phase of the two polarization channels is a good diagnostic of phase jumps in the system. Large weather-related phase changes will cancel, while phase jumps in the system will often affect the two channels differently.

Once a bad scan has been detected, it is relatively straightforward to find the offending data point (or points) that produced the large r.m.s. scatter. Most errors are associated with an antenna, hence data on all baselines associated with that antenna at the relevant times must be edited out. Interference tends to occur only on short baselines.

It is important to be consistent about the editing. For example, if the first data sample for all calibrator observations is low because data were collected before the antennas reached the source, then you should assume that the same problem occurred for all sources, even though it is not possible to detect the decrease for the weak sources. If a phase jump occurred in an antenna during an observation interspersed between two calibrator sources, then, unless there is some way to ascertain the exact time of the phase jump, it is better to delete all data associated with that antenna for the intervening observations.

6.3. Data displays

Displays using TV devices are useful for showing all of the data in a conveniently assimilated form, from which discrepant points can be recognized and discarded. The data are generally gridded into a two-dimensional matrix, with time as one axis and baseline the other. The amplitude is usually displayed as greyscale intensity or color-coded in some convenient manner. The AIPS program 'TVFLG' provides this capability.

The most useful baseline ordering is that associated with one antenna, since almost all large perturbations of the observed visibilities are antenna-based. Such an ordering might be $(V_{12}, V_{13}, \ldots, V_{1N}, V_{23}, V_{24}, \ldots, V_{2N}, \ldots, \ldots, V_{N-2,N-1}, V_{N-2,N}, V_{N-1,N})$. Errors that are associated with any antenna should be obvious from the appropriate group. For sources that are extended and that have larger visibility amplitude at short spacings, ordering the baselines by increasing separation might be more useful. For noise-dominated sources, the grey-scale or color-coding should be chosen so that points greater than 5σ can be easily recognized. It should be convenient to use a cursor on the TV display to point to discrepant data points, which can then be automatically flagged in the data base.

The editing of spectral-line data is time-consuming. A simple scheme is to scrutinize and edit the associated wide-band channel and then edit all associated spectral channels (task TVFLG in AIPS). With a display in which time is represented on one axis and spectral channel on another, individual discrepant spectral data points can be easily found and flagged (task SPFLG in AIPS).

6.4. Shadowing

When the projected spacing between two antennas is less than the diameter of the antenna, the radiation from the source to the far antenna is partially blocked by the near antenna. Although the geometrical blockage can be easily calculated, it is not generally safe to assume that the affected data can be corrected through an antenna-based multiplication, since the 'shadow' will be modified by diffraction effects. Furthermore, the blockage affects the primary beam pattern and the effective baseline length. Thus, a properly cautious procedure is to delete all data associated with the shadowed antenna. Be on the lookout for large, correlated signals, which can also occur under near-shadowing conditions.

6.5. Long-term consistency

Consistency of the data over long periods of time is difficult to determine before calibrating the data. Periodic observations of calibrator sources are important in determining long-term stability and discontinuous changes in the system. This will be discussed in Sections 7.8–7.10.

7. Final Calibration Using Radio Sources

The previous sections outlined initial calibration and editing of the visibility data, including the use of on-line monitoring. The final adjustment of the calibration is made by observations of radio sources in the sky. Calibrating the instrument using a radio source is an *ad hoc* method of calibration. It does not determine the cause of the calibration problem; it determines the complex gain of the entire system at a specific time and in a specific direction. It is therefore important to eliminate major directionally dependent calibration errors and short time-scale variations before using calibrator sources for temporal calibration. Calibrator sources are most useful for determining temporal variations over time-scales longer than about five minutes; most of these variations are associated with refraction variations over the array, caused by the troposphere and ionosphere. In this section we will investigate how to fully exploit the calibration equation 5–4.

7.1. Calibrator source properties

Radio astronomers are fortunate that there are hundreds of isolated, small-diameter radio sources with high flux density. These sources have ideal properties for a test signal: they are not significantly variable (over the time-scale of observation), they have accurately measured positions in the sky, their spectra are known and simple, they are strong enough to allow calibration in a short time, and they are isolated from nearby *confusing* sources. In short, the true visibility, $V_{ij}(t)$ is known for these sources; hence the various calibration gain terms in Equation 5–3 can be determined from the observed visibility, $\widetilde{V}_{ij}(t)$. These *test signals* are especially useful because they traverse the entire array system, from the ionosphere and troposphere, through all of the electronics, to the sampling and recording devices. They have their limitations because calibrator observations cannot, in general, be made simultaneously and at the same location in the sky as the source that is being calibrated. Thus, calibrator sources cannot calibrate gain variations that have time-scales shorter than about one minute, the typical time it takes to cycle the array between the calibrator and the source, or that have angular size-scales larger than a few degrees. A fortunate exception occurs if the target source itself is strong enough to calibrate the array (self-calibration), or there lies within the field-of-view a calibrator source (which is essentially the same thing as self-calibration).

7.2. The offset term

The baseline-based offset term $\epsilon_{ij}(t)$ is generally negligible unless a correlator is malfunctioning, or unless there is significant cross-talk between the various antenna channels. The simplest method to determine this term is to observe a part of the radio sky that contains no emission and integrate many hours in order to decrease the contribution $\eta_{ij}(t)$ from the stochastic noise sources. An hour's integration time should suffice to find correlators that have significant offset problems. Twelve-hour integration is needed to determine offsets that might cause artifacts in the images made with many tens of hours of integration time. The measured visibility on each baseline is then an estimate of the offset term. Unfortunately, at low frequency and low resolution any random location in the sky will contain emission that can be detected in only a few minutes of integration and contaminate the measurement of any offset signal.

Two methods of reducing the presence of real signal are available. The first is to calibrate the data assuming the offset term is negligible and make an image of the emission in the field. Subtract the visibility function associated with the real structure from the observed visibility data and average the residual visibility data over time for each baseline. This will give a good measure of the offsets, which can then be subtracted from the original data.

If both polarizations are observed at one frequency, the difference in the offset term can be determined. For example, if the two channels are the right-circular polarization and the left-circular polarization, their difference, which measures the circular polarization of the emission, is insignificant for most radio sources. Residual visibility in the average signal from a correlator that is much larger than that expected from noise is a measure of the difference of the offset term between the two polarizations.

These methods assume that any offset signal is reasonably constant over many hours. If the variations are short-term, they are difficult to find; but they then act more like stochastic noise, which will average out to some extent. Once the baseline-based complex offset terms are determined, they should be subtracted from the observed visibility function or flagged from the database before any other calibrations are made.

7.3. Averaging of calibration observations

In order to increase the accuracy of the computed complex gain, calibrator sources are observed for several minutes and a *suitable* average of the observed visibility is obtained. An unbiased estimate of the phase is determined by taking the vector average of the visibilities; an unbiased estimate of the amplitude is more difficult to obtain. Since the amplitude and phase fluctuations of most instrumental effects are uncorrelated, separate averaging of the amplitude and phase produce the best estimate of each. Often the amplitude instability is less than one percent, while tropospheric turbulence produces visibility phase-winds of many degrees. Vector averaging of the visibility function will produce an amplitude that is low because of the decorrelation by the phase-wind. Under these conditions, an arithmetic average of the amplitudes alone is preferable. However, if the signal-to-noise ratio for the calibrator source for each time sample is less than about five, then this simple average of the amplitude will be increased by a noise bias (see Fig. 9–2 and related discussion). An arithmetic average is generally recommended, unless the signal-to-noise ratio is very low. In this case, the vector averaging time must be kept shorter than the decorrelation time-scale.

7.4. Baseline-based calibration

The most straightforward use of calibrator observations is for determination of the complex gain $\mathcal{G}_{ij}$ for each baseline. For the sake of simplicity, assume that the calibrator source is a point source of known flux density S and known position; thus the true complex visibility amplitude is S—i.e., the amplitude is S Jy, and the phase is zero degrees. At the time of observation of the calibrator, the complex baseline-based gain is

$$\mathcal{G}_{ij}(t) = \frac{\widetilde{V}_{ij}(t)}{S} \,. \tag{5–13}$$

In words, the estimate of the gain is the observed complex visibility of the calibrator, divided by its flux density. The offset term $\epsilon_{ij}(t)$ is assumed to be negligible or already removed, and the noise $\eta_{ij}(t)$ is assumed to be negligible after proper averaging of the data in the scan. These assumptions are nearly always well-justified.

7.5. Antenna-based solutions for amplitude and phase

In Section 2.5 we noted that most data corruption occurs before the signals are correlated, so it is convenient to write the baseline-based complex gain $\mathcal{G}_{ij}(t)$ as

$$\mathcal{G}_{ij}(t) = g_i(t)g_j^*(t)g_{ij}(t)\,, \tag{5–14}$$

where

$g_i(t)$ is the antenna-based complex gain for antenna i, and
$g_{ij}(t)$ is the residual baseline-based complex gain.

Note that, if the baseline gain can be perfectly factored into a product of antenna gains, then $g_{ij} = 1$. For well-designed systems, this residual is within one percent of unity. This term is commonly (and unfortunately) called the 'closure error'.

Equation 5–14 can then be separated into an amplitude and a phase to give

$$\mathcal{A}_{ij}(t) = a_i(t)a_j(t)a_{ij}(t) \tag{5–15}$$

and

$$\Phi_{ij}(t) = \phi_i(t) - \phi_j(t) + \phi_{ij}(t)\,. \tag{5–16}$$

Let us represent the true and the observed visibilities as $V_{ij} = A_{ij}e^{i\varphi_{ij}}$ and $\widetilde{V}_{ij} = \widetilde{A}_{ij}e^{i\widetilde{\varphi}_{ij}}$, respectively. Then, for a point-source calibrator of flux density S, we have $A_{ij} = S$ and $\varphi_{ij} = 0$. The calibration equations then become

$$\widetilde{A}_{ij} = a_i a_j a_{ij} S \tag{5–17}$$

and

$$\widetilde{\varphi}_{ij} = \phi_i - \phi_j + \phi_{ij}\,, \tag{5–18}$$

which can easily be solved for a_i and ϕ_i for all N antennas, provided that the 'closure term' g_{ij} is close to unity.

Several solution methods are available. For the phase, a linear least-squares solution can be used to estimate the antenna-based phases. Because the equations depend on differences between the antenna-based phases, one antenna phase is arbitrary, and is normally taken as zero. The antenna gains can be solved for by linearizing the gain equation through use of logarithms, then utilizing least-squares. This method, unfortunately, is susceptible to a biasing effect, due to noise—so that weak calibrators cannot be used to obtain reliable estimates of the gain amplitudes. A better approach is to solve the complex equation simultaneously for the amplitude and phase.

An estimate of the residual baseline-based error, often called the *closure phase* (unfortunately not the same as the structure-dependent closure phase), can be obtained from

$$\phi_{ij} = \widetilde{\varphi}_{ij} - \phi_i + \phi_j\,. \tag{5–19}$$

This quantity should be less than one degree for a well-designed system. Similarly, the residual amplitude can be determined from

$$a_{ij} = \frac{\widetilde{A}_{ij}}{a_i a_j S} \tag{5–20}$$

and should not exceed 1.01 for a well-designed system.

There are three reasons why antenna-based calibration values are preferred over baseline-based ones. First, most variations in the instrument are related to particular antennas, whether they arise from the medium above the antenna or the electronic components. The errors due to a well-designed correlator are always vastly smaller.

Secondly, since the number of baselines in an array of N elements $(N > 4)$ is far larger than the number of elements, the computer capacity needed to store the calibration parameters and to apply them to the observed visibilities is reduced using antenna-based relations. Even if the baseline-based errors are significant, they are often constant with time—so a single determination of these errors can be used globally, to correct a large set of data.

Finally, antenna-based calibrations can be determined without the full set of baseline data. This is extremely important when using calibrators which are partially resolved, or confused by nearby sources. Such calibrators are often adequately approximated by a point source over a limited range of baseline lengths. Provided that all correlations between antennas are obtained, good calibration solutions can still be obtained in spite of stringent restrictions in the allowed range of spacings.

7.6. Imperfect calibrators

Calibrators are, in general, not point sources. For high-resolution arrays, as in VLBI, all bright sources are resolved. At lower resolution and lower frequencies, many VLA calibrators can be considered point sources, but with extended emission present and with susceptibility to confusion from nearby sources. Partially resolved sources are still useful, provided the visibility is nearly enough constant over a sufficient range in baseline spacing. A good example of this is shown in Lecture 13.

If the calibrator source has a complex structure and cannot be approximated by a point source for any reasonable range of baselines, then self-calibration techniques are necessary to derive the antenna-based complex gain. The details of this process are given in Lecture 10. In that lecture the main impetus is to derive a good approximation of the source brightness distribution. However, the antenna-based complex gains must also be derived and it is these which are required for the calibration of the instrument. The calibration source should not be too complex, in order to avoid the ambiguities of determining both the structure and the complex gain.

7.7. Directional calibration—astrometry

Although an important reason for frequent calibration observations is to determine the temporal changes of the array (particularly those caused by refraction in the troposphere and ionosphere), there is always a residual directional dependence of both the amplitude and phase of the gain. The effect of this dependence can be minimized if the calibrator source is close to the field that is being imaged, but it is best to determine the directional dependence as accurately as possible.

In most array operations, special calibrator observations are made to determine the directional dependencies, which are generally stable over periods of days. This often is done by observatory support staff, rather than observers themselves. Once these corrections have been determined, observers are free to concentrate their own calibration efforts on the problem of temporal variations (i.e., excluding directional effects). The typical observational scheme to determine the directional dependencies is to observe a large number of good-quality calibrator sources well-distributed over the sky, and packed into as short a period of time as possible—typically, about twenty observations in one hour. Observations at night and during clear, windless weather conditions, when tropospheric refraction variations are minimal, are the most productive. The sky coverage and the type of calibrator sources depend on the parameters that are to be calibrated. When carried to its extreme, this type of calibration is called *astrometry* and the following parameters can be determined:

1. Accurate locations of the antennas;
2. Antenna structure parameters (axis intersection displacement, for example);
3. More accurate positions of the calibrator sources;
4. Earth rotation (clock error);
5. Earth orientation (polar motion);
6. Zenith delay of the troposphere or the ionosphere;
7. Amplitude versus elevation dependence; and
8. Apparent displacement by the solar gravitational field.

When this type of observation is made periodically, changes in the above parameters can be monitored. Motion of the location of antenna which are on different tectonic plates are important geophysical data. Better values for the Earth precession, nutation, polar-motion traces, tidal motions, Earth-rotation changes, etc. can be determined.

Many of the above effects are of the order of milli-arcseconds and require very high-resolution arrays—particularly VLBI arrays, with antennas separated by many thousands of kilometers. Because these effects have similar directional dependencies, many observations of calibrators all over the sky are needed; complicated analysis methods are used, based on models similar to Equation 5–3, and embodying the geometric dependence of each relevant effect. When VLBI techniques are used, the delay (called *group delay* by VLBI'ers) is used rather than the phase delay, and a sophisticated form of the delay calibration of Section 3.2 is applied. The NASA software program known as 'CALC' makes initial estimates of these parameters, and the program 'SOLVE' then determines refined values from a set of antenna-based phases or group delays. See Lecture 23 for more details.

7.8. Checking the calibration: closure errors

An important step in calibrating and editing the data is to check the validity of the assumption that the complex gain of the observations can be approximated by antenna-based solutions—for each antenna, frequency and polarization. The soundness of this assumption is determined by the magnitude of the closure error. Some examples and suggested remedies are discussed below.

The closure errors are defined in Equations 5–19 and 5–20. They are, in essence, those residual amplitudes and phases for each baseline, averaged over the scan, which—if applied to the true visibility function after application of the best antenna-based gain—would reproduce the observed visibility function. For most arrays, amplitude closure errors should be less than 2%, and phase closure errors should be less than 2°.

If several baselines show significant, constant closure errors using different calibrator sources, then g_{ij} is different from unity and closure correction should be applied to all of the visibility measurements. If all of the large closure errors are associated with baselines to one antenna, the data from that antenna are suspect and should be closely scrutinized.

If closure errors are mainly associated with short baselines or with long baselines, then the calibrator source may not be sufficiently point-like. Structure in a radio calibrator can be confirmed if both polarizations at the same frequency show similar closure errors for the same baselines. Restricting the range of baselines can improve the solution quality. At last resort, self-calibration techniques that determine both the source structure and the antenna-based gain solution may have to be used.

Occasional large, anomalous closure errors on some baselines are indicative of interference or very discrepant data within a scan. Such data can usually be found by scrutiny of the individual points in the scan, for the relevant antennas or baselines. Interference spikes, extreme instability, contaminating signal from the Sun, or radar can be the cause.

Closure errors can be recognized only in calibrator observations, because their effects are subtle on a source with any degree of complexity. After the data problems have been determined, careful judgement must be used in applying closure error corrections derived from calibrator observations. For example, if just one antenna is producing large closure errors for many baselines for all calibrator data, it is best to flag all that antenna's data during the period of misbehavior. Under these conditions, it is likely the bad data affected the previously determined antenna gains, so it is wise to recalibrate.

Closure errors of a few percent are present in most observations at the VLA. They are caused by a variety of subtle problems: timing problems in the correlator, amplitude and phase variations over the frequency band, delay errors, and non-quadrature of the real and imaginary parts of the measured visibility function (e.g., Thompson 1980). All of these effects are stable over days. These closure errors produce artifacts in images at a level of about 5000 : 1. If you wish to image a very strong source then it is best to calibrate these closure errors as well as possible. There are several very bright point sources (3C 84 for example) that produce sufficient signal-to-noise ratio to allow determination of closure errors and their stability. Once determined, these closure corrections should be applied to the data before obtaining antenna-based calibrations or running the self-calibration algorithm.

7.9. Checking the calibration: amplitude stability

During observing, many calibrator observations are made; afterward, the observer generates a table of the amplitude calibration of each antenna, for all frequencies and polarizations. If the array is operating well, the amplitude for any antenna should not vary over the observation period by more than a few percent; also, the amplitude ratio between antennas should be in the range 0.8 to 1.2. Problems frequently arise, however.

The flux density of some or all of the calibrators may be in error. Nearly all good point calibrators are variable, especially at high frequencies. More accurate relative flux densities for all of one's calibrators can be obtained from the ratio of the amplitude calibration from any or all of the antennas for the different sources. For example, if all of the antenna amplitude solutions for source A are 1.2 times greater than those for source B, then one can assume that the flux density of source A is a factor $\sqrt{1.2}$ less than that of source B. The absolute flux density scale for the observations cannot be measured directly. Some radio sources have accurately-measured flux densities over a wide range of frequencies; one or more of these sources should be observed a few times during the run (3C 286, 3C 48, Planets, Planetary Nebulae).

Jumps or changes in the amplitude calibration for some antennas is an indication of real instabilities. Whether to flag these data depends on the type of experiment and the severity of the instability. If there is correlation of the instabilities in the two polarizations, then the problem may be due to shadowing or mis-pointing of the telescopes. The delay-zero may also be in error. If all of the antennas follow the same changes in amplitude, then residual gain vs. elevation problems (associated with sky absorption, antenna inefficiency, or ground pickup) may be occurring.

7.10. Checking the calibration: phase stability

For most observations the phase stability is determined by the tropospheric and ionospheric refraction turbulence over the array, the most important component being the water-vapor distribution. How well frequent calibrations correct for this phase error in a nearby source will be discussed in Section 8. If both polarizations at one frequency are simultaneously observed, their phase difference should cancel all refraction variations; this is a good diagnostic of the system phase behavior. The difference should not vary by more than 5° over twelve hours. Many system phase jumps are different at the two polarizations and can be seen in the difference. If the antennas are separated by more than 100 km and the ionospheric refraction is large, then differential Faraday rotation of the two circular polarizations may be significant. See Section 9.3.

There are several tests of the phase variations that will indicate whether they are likely caused by weather. First, the fluctuations should roughly scale with frequency if they are tropospheric or scale with wavelength if they are ionospheric. Secondly, the fluctuations will scale with baseline length, b. For baselines less than 2 km the scaling goes as $b^{0.8}$ which is indicative of Kolmogorov turbulence. For baseline between 2 km and 100 km the scaling goes as $b^{0.3}$. For longer baselines the variations do not increase much. These relations are only approximate (Sramek 1983). Of course, the general tendency of the phase fluctuations should be correlated with the weather. Dry, windless days should be stable; periods of afternoon thunderstorms should be dreadful; nights are better than days; solar minimum is better than solar maximum.

Look for phase drifts in the antenna phase solutions over a period of several hours. If phase drifts are noticeable in one or a few antennas it is likely that the assumed antenna positions are in error. Check with the array operator in order to determine the status of the antenna position calibration. If many of the antennas show phase drifts that scale with baseline length from the reference antenna, then the source position is probably in error. A better position can be obtained from imaging or by use of Equation 5–3. If the phase drift occurs for all of the sources, then the array time standard may be in error.

7.11. Calibrator choice

In general, the closer the calibrator is to the source the better: direction-dependent gain terms then are less important. Some improvement in the calibration of temporal fluctuations caused by the troposphere occurs with a calibrator only a few degrees away from the source, but most of the benefit is lost because of the non-simultaneity of the calibrator and source observations. The most conservative calibration scheme is to observe in the sequence calibrator–source–calibrator. If you are particularly concerned about direction-dependent errors, then the use of several calibrators surrounding the source (or two calibrators collinear with the source) is recommended. These extreme configurations are needed for VLBI observations if phase coherence is desired. The faster the cycling time between sources, the better. However, most of the time will be used for calibration, and this may make the signal-to-noise ratio for the source too low. See Lectures 22, 28 and 29 for guidelines specific to a given frequency and array.

8. Application of the Calibration Values

Depending on their origin, the calibration values are applied to the measured visibility function with different time-scales and in different ways. After a major reconfiguration of the array or a substantial change in the electronics, approximate parameters are introduced into the system (antenna locations, delay zeros, antenna pointing parameters)—generally in the on-line computer that controls the telescope, collects the data, and records it on tape. Since these values may be in serious error and contaminate the data quality, initial calibrations are required, as described in Section 3. The relevant parameters must be changed in the on-line computer before serious observing is done. Other direct calibrations (Sec. 4) are applied as the data are collected—for example, the path length monitoring of the local oscillator and the automatic level correction.

8.1. Gain tables

Some calibrations are uncertain, so they should not be applied directly to the data unless in a form which can be undone. For example, correction of the path length due to the tropospheric delay, as determined by ground-based weather monitoring, is uncertain and may be revised at a later time. If these calibrations are relatively slowly changing they can be stored in what is called a *gain table*. This is a tabular array of numbers that contains the antenna-based amplitude and phase calibration values (or baseline-based values) at convenient intervals of time. Whether each calibration step has its own gain table or all of them are lumped into one table depends on the software system. But, it is important to be able to reproduce the values associated with any calibration—in order to undo its effect, if necessary. At the VLA, the practice is to have gain table entries corresponding to the beginning and end of each scan, and entries at ten-minute intervals within long scans. For VLBI observations, gain tables are usually produced at shorter intervals. The changes in the amplitude between gain table entries should not exceed more than about 25%; the changes in phase should not exceed more than about 30°—otherwise interpolation of the gain entries to intermediate times will be too much in error.

8.2. Interpolation within the gain table

The calibration of the array using radio sources, discussed in Section 7, produces a table of antenna-based amplitude and phase calibrations for each antenna, frequency, and polarization at the mean time of the calibration scan. These values are separated by the time between calibrator observations, typically ten to forty minutes. How should these values be interpolated for times between calibrator observations? If the calibration variations are small, it matters little what interpolation scheme is used to transfer the values into the gain table. Similarly, if the calibration variations are very large, it also makes little difference, since the detailed fluctuations in the calibrator may not be highly correlated with those in the source. The recommendation here is to use a relatively long averaging time of the calibration values (running means over periods that include about three observations) and apply these smooth values to all the data. This scheme will calibrate the longer-term variations, and the residual errors left in the calibrator observations themselves will be a reasonable indication of the residual errors in the source observations. Regardless of the interpolation scheme, the amplitude and the phase should be interpolated separately; not the real and imaginary parts of the antenna-based gain.

8.3. Large phase-fluctuations

On long baselines and in periods of poor weather there are large differences in the antenna-based phase between successive calibrator observations. Unless the phase change is consistent with a linear drift (look for a drift during the calibrator scan), it may be difficult to (sensibly) interpolate between the calibrator observations. Lobe ambiguities will occur if the phase change between two successive calibrator observations is near 180°. With this ambiguity, interpolation may give entirely the wrong result. So, in the extreme case we say that the array is *incoherent* over the time-scale of the calibration gaps. It may be coherent for shorter time-scales and for shorter baselines.

One method of sorting out possible lobe ambiguities is to determine the magnitude and direction of the phase difference between the two calibrator observations at short baselines. Since phase fluctuations are coherent over many kilometers, lobe ambiguities of long baselines may be resolved by the phase change at the shorter baselines. Similarly, if observations are made simultaneously at another frequency, the frequency scaling of tropospheric fluctuations from the lower frequency may be an aid. Fluctuations from the ionosphere can be handled somewhat differently and are discussed in Section 10.

9. Polarization Calibration

As shown in Lecture 1, the complete state of the radiation field is most conveniently described by the superposition of two orthogonal polarizations. The complete spatial coherence of this vector field is described by four correlations. For example, if the dual-polarization feeds measure the right circular polarization (R) and the left circular polarization (L), then the four visibility functions are $R \star R$, $L \star L$, $R \star L$, and L ⋆ R. The previous discussion concerned the calibration of the parallel hand correlations, $R \star R$ and $L \star L$. With these two visibility functions, the complex gains of the two polarization channels can be calibrated. If the dual-polarization channels were precisely orthogonal (that is, if the voltage signals from the channels were precisely proportional to the orthogonal electric fields), then the cross-hands would also be calibrated, except for any phase difference between them and for possible correlator-based problems. In practice, however, this proportionality is not obtained, and further calibrations are necessary. This section describes the additional calibrations needed for the crossed-hand correlations, $L \star R$ and $R \star L$. For notational simplicity, we have assumed the parallel-hand calibration has been accomplished.

9.1. Stokes parameters

Section 5.2 of Lecture 1 has described the polarization matrix and its relation to the Stokes' parameters, I, Q, U, and V. These are four real numbers describing the polarization state of radiation from a given region of sky, for a given frequency. Note that Q, U and V can be negative. It is quite correct to consider these as four images which completely characterize the polarization state of the radiation from the sky. Each of these has an associated visibility function, which we denote V_I, V_Q, V_U, and V_V. These cannot be measured individually, but appear as linear combinations of the four correlations produced by the interferometer. From these combinations, the four visibility functions can be obtained. With the use of R and L polarized feeds, the four visibility functions are

$$\begin{aligned} V[R \star R] &= V_I + V_V\,, \\ V[L \star L] &= V_I - V_V\,, \\ V[R \star L] &= (V_Q + iV_U)\ e^{-2i\chi}\,, \\ V[L \star R] &= (V_Q - iV_U)\ e^{2i\chi}\,, \end{aligned} \tag{5–21}$$

where χ is the parallactic angle, which determines the orientation of the feed with respect to the sky. Other combinations are given by Thompson *et al.* (1986, Sec. 4.9).

9.2. Leakage terms

The feeds are not exactly orthogonal; that is, the antenna-feed combination causes a small amount of right-circular polarization to show up in the left channel, and *vice versa*. This 'leakage' is defined in the following way: If the provided voltages are v_R and v_L from the RCP and LCP channels, and the true electric fields are E_R and E_L, then $v_R = E_R e^{-i\chi} + D_R E_L e^{i\chi}$ and $v_L = E_L e^{i\chi} + D_L E_R e^{-i\chi}$. Note that the voltages are considered to be complex, corresponding to what is called the 'analytic signal' in the engineering literature. The indicated multiplications can be carried out, and assuming the leakage terms, and the circular and linear terms, are small so that products of them may be ignored, the four correlations can be expressed as (Bignell & Napier 1978):

$$\begin{aligned} V[R\star R] &= V_I + V_V\,, \\ V[L\star L] &= V_I - V_V\,, \\ V[R\star L] &= e^{-2i\chi}(V_Q + iV_U) + (D_{R1} + D^*_{L2})V_I\,, \\ V[L\star R] &= e^{2i\chi}(V_Q - iV_U) + (D_{L1} + D^*_{R2})V_I\,. \end{aligned} \tag{5–22}$$

For notational clarity, we have left off a phase rotation $e^{\pm i(\phi_R - \phi_L)}$ that should multiply the final two equations. This term is due to the arbitrary phase difference that occurs between the two orthogonally polarized channels. This phase difference is easily determined from observations of a strongly polarized source of known position angle, as described in the next section. χ is the relative angle of the feed orientation with the sky. For an equatorially mounted antenna, the parallactic angle is constant unless the feeds are purposely rotated. For an alt–azimuth mounted antenna, the parallactic angle is a calculable function of time and position: $\tan\chi = \cos\phi\sin h/(\sin\phi\cos\delta - \cos\phi\sin\delta\cos h)$.

The leakage terms D can be calculated by observing an unpolarized calibrator point source. In this case, the observed visibility in the cross-hand channels is a sum of two leakage terms. Using all of the baselines, the two leakage terms per antenna can be obtained. In general, the calibrator polarization is not zero, and must be included in the solution. For an alt–az antenna, both the source polarization and antenna polarization can be simultaneously calculated, since the change $\Delta\chi$ in parallactic angle causes the source and antenna polarizations to differentially rotate. The varying sum can be used to determine each term. The accuracy of this separation is dependent upon the strength of the source and the range of parallactic angle.

9.3. Phase difference between polarization channels

The normal parallel-hand phase calibration calibrates the phase of one polarization channel of each antenna with respect to the phase of that same channel of the reference antenna. This is done independently for each polarization channel. In general, there will then remain a phase difference *between* the two polarization channels. The parallel hand calibration ensures that this difference will be the same for each antenna, and equal to that of the reference antenna. Since the crossed-hand phase carries information on the visibilities of Stokes' parameters Q and U, it is clear that this instrumental phase difference must be removed.

The phase difference of the reference antenna can be measured by observing a strongly linearly polarized radio source with known parameters $Q + iU$. Since we need only the phase difference for one antenna, only one baseline is needed; a choice of one of the shortest baselines will permit the use of highly polarized calibrators, which tend to be somewhat resolved.

In the ionosphere, the presence of a magnetic field causes the refraction to be different for right- and left-circularly polarized radio waves. This produces a rotation of the plane of the linearly-polarized signal and, equivalently, a phase rotation between the R polarization channel and the L polarization channel for the reference antenna. Since the phase calibration of the other antennas has been tied to the reference antenna, the difference is uniform over the array. If this differential Faraday rotation changes over the array, as it will for VLBI arrays, then the R and L phase calibrations for these antennas will differ. During periods of high solar activity the phase rotation can be as large as 20 rad m^{-2}, which produces an angle of 50° at 1.4 GHz. The angle scales with the square of the wavelength.

Ionospheric models are available to predict the density of electrons and the strength of the magnetic field, from which estimates of the differential Faraday rotation can be made. Abrupt changes occur at sunrise and sunset, and significant changes occur during the day with time-scales of about one hour. For best results, at least once an hour observe a highly polarized calibrator source not more than about 20° from the source.

10. Additional Topics

A *potpourri* of topics follows in this section. Some deal with fairly subtle and unusual calibration practices but are worth mentioning. Others deal with calibration of the entire field-of-view of the observations. Nearly all of the discussion here and elsewhere in the lecture deals only with the central region of the primary beam.

10.1. The antenna primary beam

In Section 3.1 the calibration of the antenna pointing was discussed. Since the visibility function measures the spatial coherence function of a fictitious source $\mathcal{A}(l,m)I(l,m)$ (see Eq. 5–1), the effect of the primary beam sensitivity, $\mathcal{A}(l,m)$, must be removed from the image. Although a good theoretical estimate of $\mathcal{A}$ can be made of the central region, above a relative sensitivity of 20%, direct observations are needed to determine $\mathcal{A}$ in the regions $< 20\%$ and in the sidelobes. Observations are made of a very strong calibrator. Assuming all of the antennas are identical, you choose a convenient antenna pair, point one directly at the source and mis-point the other over the area of sky in which you want to determine the primary beam sensitivity. You take occasional observations with the second telescope pointed directly at the source, in order to determine variations of the antenna amplitude and phase with time. The complex ratio of the visibility measured at any desired location in the sky to visibility measured on-axis is the normalized reception *voltage* pattern. If two polarizations are observed, this procedure should be done for all four polarization correlations.

An interesting sidelight is that the Fourier transform of the complex relative visibility function, as a function of distance from the phase center, gives the distribution of the electric field on the aperture plane of the antenna. Departures of the phase from zero can be interpreted as surface deviations from a true paraboloid and thus measure the accuracy of the surface. This technique is commonly used for surface measurements of large telescopes, and is called holography (see Lectures 3 and 28).

The absolute value of the voltage pattern is $\mathcal{A}(l,m)$. The image made from the Fourier transform of the calibrated visibility function must be corrected for the primary beam sensitivity in order to determine the distribution of intensity in the sky. This correction becomes crucial when very large sources that cover several primary beam areas are to be combined into one image (see Lecture 20), when comparisons at widely different frequencies are made, and when the polarization properties of large sources are of interest.

Several other problems associated with the antenna pointing and finite size occur and will be briefly mentioned:

1. Beam-squint between polarizations: For off-axis feed systems like those of the VLA and VLBA, there is a slight displacement of about 0.05 FWHP between the RCP and LCP electrical pointing centers. For accurate images of very extended sources, the primary beam correction of each must be done separately. The image difference between the two parallel hands will often show false circularly polarized sources where the two primary beam corrections differ.

2. Polarization characteristics across the beam: The calibration of the leakage terms D, described in Section 9 and also in Lecture 6, can be made at each location in the primary beam. It is found that the D's substantially change outside of the 50% sensitivity region of the beam. Thus the relative linear combinations of the four Stokes parameters are mixed-up across the beam, invalidating the basic synthesis equation. And thus, outside of this area, it is difficult to reliably measure degrees of linear polarization lower than about 5%.

3. Alt–azimuth antenna mount: For the VLA and the VLBA the alt–azimuth mounting produces a rotation, measured by the parallactic angle χ, of the antenna relative to the sky. Unless the primary beam response is perfectly circular (i.e., identical along all radii), extended observations will cause an effective change in the primary beam sensitivity across the source. Such non-circularity is more pronounced for the cross-hand correlations than for the parallel-handed. A (tedious) solution is to image only that part of the data in which the parallactic angle has not changed by more than, say, 10°, correct for the instantaneous primary beam sensitivity, repeat, and then sum up all the corrected intensity distributions.

4. Different antenna in the array: In general, the synthesis properties of an array with more than two elements is invalidated if the antennas are dissimilar. This is because the effective primary beam sensitivity is different for different baselines and the sampled coherence function is not uniquely associated with one intensity distribution. For two-element arrays, the product of the voltage primary beam patterns applies. Many VLBI arrays are composed of widely different diameter antennas; since all of the sources under study are extremely small in angular size, they are all virtually at the beam center, so the details of the primary beam patterns are irrelevant.

10.2. Bandwidth smearing

When observing over large instantaneous bandwidths, distortions occur in parts of the image far removed from the phase center. This effect, *chromatic aberration*, is discussed in detail in Lectures 17 and 18. As a rule-of-thumb, chromatic aberration becomes significant when the offset, measured in units of synthesized beamwidths, and multiplied by the fractional bandwidth, is of order unity. It produces a radial smearing whose shape is the effective bandpass convolved with the source structure. Deconvolution techniques, generalized to recognize this smearing, can be used to obtain an estimate of the intensity distribution.

The smearing can be avoided only by separating the observations into a set of narrower-bandwidth channels by using filters, or by cross-correlation techniques in which many delay centers are used (see Sec. 5).

10.3. Dual-frequency ionospheric calibration

Over most of the radio frequency spectrum, the ionospheric refraction varies as the wavelength squared, so the effect of the ionosphere on the visibility phase is proportional to the wavelength. If a point source is observed, or the structure of the source is known or can be assumed to be invariant between two frequencies, the effects of the ionosphere on the phase can be removed (Fomalont & Sramek 1977). Given two observations of the source phase, ϕ_1 and ϕ_2 at frequencies ν_1 and ν_2, the corrected phase ϕ_c, which would have been measured in the absence of the ionosphere, is:

$$\phi_c = \frac{\nu_2\phi_2 - \nu_1\phi_1}{\nu_2 - \nu_1}\,. \tag{5–23}$$

The corrected phase is the sum of the true visibility phase and an instrumental phase. The latter can be determined from an observation of any calibrator. This technique is used for geodetic VLBI, with the delay replacing the phase. Common dual-frequency pairs are (2.3 GHz, 8.1 GHz) and (1.3 GHz, 1.6 GHz).

10.4. Relative *vs.* absolute calibration

If the radio source is strong and the signal-to-noise ratio of the visibility is greater than about five (for each integration time and for each baseline), then self-calibration techniques will likely succeed. In principle, there is no need to calibrate the data at all. It is advisable to observe a very strong source to determine the closure errors, and blank sky to determine the offset terms. If polarization information is required, proper calibration of the leakage terms is needed, using an unpolarized source and observations of the polarized source to calculate the differential Faraday rotation. An observation of a flux density standard is necessary to determine the flux density scale.

Alternating the observations between the source and a nearby calibrator aids in removing some of the tropospheric refraction errors and often produces an image with dynamic range $> 100:1$. The position of the source then is known with respect to that of the calibrator, but with some residual error caused by any long-term directional dependencies of the phase. Even if self-calibration techniques will be useful, this normal calibration will produce a good first image that may be then used for further self-calibration. When one uses the no-calibration scheme described in the last paragraph, a good first estimate of the source structure is not available: this is the usual case in VLBI observations. If the source is very complicated and the (u, v) coverage is poor, then ambiguities in the self-calibration process may become important, especially in the absence of a good starting model of the source. Relative motion of the source, or of parts of the source, over long periods of time can be determined if the same calibrator is used for all observations. Registration of images made at different frequencies, at different times, and with different configurations are made easier by using the same calibrator source. If the radio field contains many radio sources, then the images at different epochs can be registered by aligning the background sources in the final images, in order to determine relative motion of the source of interest.

True absolute measurements of radio source positions or antenna positions require careful calibration of the directional dependence of the phase, as described in Section 7.7. The origin of right ascension is a fundamental problem. Full use of Equation 5–3, with many additional terms, is needed. Often, corrections are made to the zenith path delay as a function of time.

References

Bignell, R. C. & Napier, P. J. 1978, VLA Test Memorandum No. 125, NRAO.

Brosche, P., Wade, C. M., & Hjellming, R. M. 1973, *ApJ*, 183, 805–818.

Carter, W. E., Robertson, D. S., & MacKay, J. R. 1985, *J. Geophys. Res.*, 90, 4577–4587.

Clark, B. G. 1973a, VLA Computer Memorandum No. 104, NRAO.

Clark, B. G. 1973b, VLA Computer Memorandum No. 105, NRAO.

Clark, B. G. 1982a, VLA Scientific Memorandum No. 145, NRAO.

Dewdney, P. 1987, VLA Test Memorandum No. 148, NRAO.

Fomalont, E. B. & Sramek, R. A. 1977, *Comments on Astrophys.*, 7, 19–33.

Sramek, R. A. 1983, VLA Test Memorandum No. 143, NRAO.

Thompson, A. R. 1980, VLA Electronics Memorandum No. 192, NRAO.

Thompson, A. R., Moran, J. M., & Swenson, G. W., Jr. 1986, *Interferometry and Synthesis in Radio Astronomy*, John Wiley & Sons, New York.

Uson, J. 1986, VLA Scientific Memorandum No. 157, NRAO.

Wade, C. M. 1970, *ApJ*, 162, 381–390.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

6. Polarization in Interferometry

W. D. Cotton
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A.

Abstract. This lecture covers the calibration and imaging of interferometric measurements of polarized radiation. The topics included are: basic concepts of polarized radiation, the instrumental response to polarized emission, calibration to remove the corrupting effects of the atmosphere and the instrument, and imaging in polarized light. Both circularly and linearly polarized feeds are discussed.

1. Introduction

Most of the non-thermal processes that produce radio frequency emission in astronomical sources are at least partially polarized. In addition, magnetized plasma along the line of sight to the source can further modify the polarization state of the emission. Thus, polarized emission is an important astrophysical diagnostic of the physical conditions in and in front of radio sources.

For several decades, radio interferometers have routinely measured the polarization of the emission (Morris et al. 1964; Conway & Kronberg 1969, Fomalont & Wright 1974; Thompson, Moran & Swenson 1986; Cotton 1993). As will be shown later, interferometers always measure some aspect of the polarized emission. There are a variety of effects that can either corrupt the measurements of polarized emission or complicate its interpretation. Many of these effects are the results of the atmosphere and the instrument itself and may be determined and corrected.

The details of polarization measurements depend strongly on the type of detector used for the radiation. The major types of detectors, or “feeds”, couple either orthogonal circularly or linearly polarized emission into the telescope electronics. The type of feed used in an interferometer profoundly affects the calibration procedures.

The following sections will present some basic concepts of polarized electromagnetic radiation, describe the response of an element and an interferometer to polarized emission, discuss various potential corruptions of the polarized signals and their cures, and finally discuss the imaging of polarized emission. This discussion will attempt to be fairly general and will not include the details of the procedures in any particular software implementation.

1.1. Polarization Concepts

Electromagnetic radiation can be thought of as propagating sets of oscillating electric and magnetic vectors. As radio telescopes are sensitive only to the electric field, the following discussion will ignore the magnetic field. Figure 6–1a illustrates the concept of light as an oscillating electric field; the curved line represents the trace of the tip of the electric vector. For a given wave, this oscillation can take a variety of forms, several of which are illustrated in Figures 6–1b–d, which show the motion of the tip of the electric vector as viewed along the direction of propagation of the wave. The polarization state of the wave is simply the shape of the trace of the tip of the electric vector in this projection.

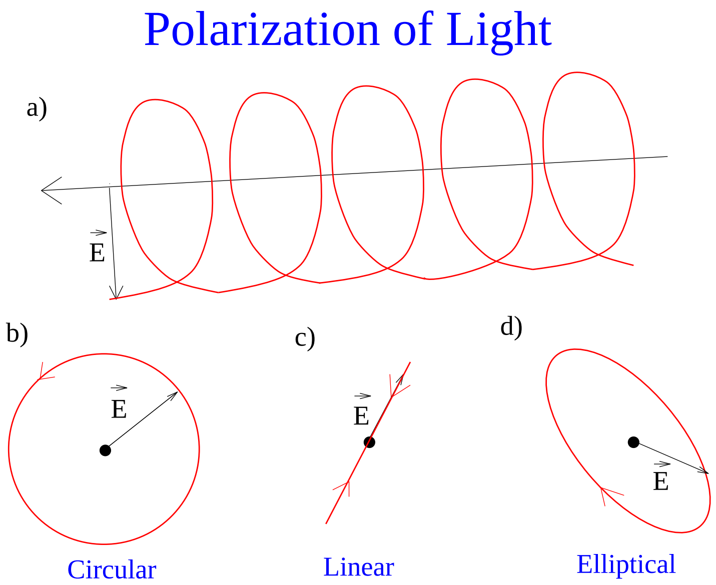


Figure 6–1. a) Illustrates the concept of electro-magnetic radiation as a propagating, oscillating electric vector.
b) End on view of the trace of the tip of the electric vector for a circularly polarized wave.
c) Like b) but for a linearly polarized wave.
d) Like b) but for an elliptically polarized wave.

Figure 6–1b illustrates circular polarization in which the E vector traces a circle; the two possible directions are called right and left circular polarizations. The E vector shown in Figure 6–1c is confined to a line and is referred to as linear polarization; the orientation of the line is arbitrary. Finally, combinations of circular and linear polarization are possible, giving rise to elliptical polarization as illustrated in 6–1d. For a detailed discussion of the polarization properties of light, see Born & Wolf (1975).

Radio sources are generally incoherent emitters of radiation — which complicates the issue of polarization, as the many different wave packets making up the light from the source may have independent polarization states. Thus, the polarization of the emission becomes a statistical measurement. If there are equal numbers of waves in a given polarization state and in the orthogonal polarization state, then the light is said to be unpolarized. Here right and left circular are “orthogonal”, as are linear polarizations with angles separated by 90°. Thus, a source with “no circular polarization” actually has equal mounts of right and left circular polarization. A source is said to be partially polarized if any polarization is more common than its orthogonal state. A further concept is that of the total intensity which is the sum of all polarization states. The polarization of a source is described as the fraction of the total intensity of the excess of the given state over the orthogonal state. For linear polarization, the orientation on the sky of the dominant polarization state is called the “polarization angle”.

A single phase sensitive detector of electro-magnetic radiation can sample only a single polarization state. However, the signal is fully described in terms of two orthogonal polarization states so detectors coupled to two orthogonal polarization are sufficient to measure the polarization state of the signal.

The polarization of a partially polarized source can be described in terms of the Stokes parameters (Stokes 1852) I, Q, U, and V. I is the total intensity, Q and U describe the linear polarization (Q points north and U is 45° towards east), and V is the circular polarization (positive is right circularly polarized). (Note: I is always positive but the other Stokes parameters can have either sign.) The polarization angle, or apparent orientation of the projected E vectors on the sky, measured from north towards east, is related to the Stokes parameters by

$$\Phi = \frac{1}{2}\tan^{-1}\frac{U}{Q} \tag{6–1}$$

The fractional linear polarization is

$$m = \frac{\sqrt{Q^2 + U^2}}{I}, \tag{6–2}$$

and the fractional circular polarization is

$$v = \frac{|V|}{I}. \tag{6–3}$$

In this lecture, the Stokes parameters will refer to visibility measures which are the same as the brightness distribution only for point sources. The Stokes parameter brightness distributions of extended sources are the Fourier transform of the Stokes parameter visibility function. For a more rigorous discussion of the Stokes parameter representation of partially polarized radiation, see equations 1-16, 1-17, 5-21, and 5-22, or Chandrasekhar (1960) and Kraus (1966).

1.2. Polarization Measurements Using an Interferometer

Polarization measurements are generally made using a pair of feeds on each interferometer element; usually these are sensitive to orthogonal circular or linear polarizations. (For technical reasons it is not possible to construct a feed sensitive to total intensity.) If polarization measurements are desired, then all four combinations of feeds on all baselines are cross-correlated. Even for total intensity measurements, both of the correlations of the parallel polarization feeds are required (although if the system uses circularly polarized feeds and is observing sources with no circular polarization, then one correlation of parallel feeds is sufficient). The relationship between the measured correlations and the Stokes parameters depends on the polarization type of the feeds and is the subject of Section 2.

1.3. Errors and Calibration

There are a number of effects that can modify or corrupt the polarized signals. These need to be evaluated and the derived corrections must be applied to the data before it can be used. Polarized signals can be modified in passage through the atmosphere. In particular, the ionosphere is a magnetized plasma capable of Faraday rotation, i.e. the rotation of the polarization angle of linear polarization.

The antenna feeds will not have an ideal response, which will cause an "instrumental" contribution to the measured polarized signal even for a completely unpolarized source. The phase relation of the signals detected by the two feeds of any given interferometer element can be disturbed both by ionospheric Faraday rotation and by the interferometer electronics. Interferometers using circularly polarized feeds do not directly measure the polarization angle of a source, so observations of a calibrator of known polarization angle are needed. Similarly, interferometers using linearly polarized feeds may have unknown orientation errors and may have trouble separating calibrator and instrumental polarization. This is especially true if the antennas have equatorial mounts, in which case, observations of a source of known polarization may therefore be required. The process of determining and correcting these errors in the data is known as calibration. The details of calibration depend on the feed polarization type and are the subject of Section 3.

2. Instrumental Response to a Polarized Signal

This section considers the response of a telescope to a polarized signal. A parameterized model of the response of each feed is needed in order to quantitatively characterize it. The parameters of this model can be determined and used to correct the data. There are several formalisms for describing the polarization response of radio telescopes; one is to describe the feed in terms of the polarization state to which it is sensitive, i.e. to the polarization that it would transmit if the antenna were used as a transmitter rather than a receiver. In this formalism, each feed is parameterized by the ellipticity and orientation of the polarization to which it is sensitive; see Fomalont and Wright (1974) or Cotton (1993) for a more detailed discussion.

This lecture will use an alternate description of the polarization response in which the feed is assumed perfectly coupled to the nominal polarization, with the addition of a complex factor times the orthogonal polarization. This is called the "Leakage" or "D–term" model. Leppänen (1995) discusses the equivalence of these two models of the feed response.

Schwab (1979) has pointed out that the response of an interferometer can be factorized into antenna based components which can be further factorized into discrete contributions arising from a number of effects. These individual contributions to the instrumental response are expressed in terms of "Jones matrices". A more detailed discussion of the Jones matrices can be found in Lecture 32, in Hamaker, Bregman & Sault (1996) and in Sault, Hamaker & Bregman (1996).

2.1. Feed Response

The response of antenna i with orthogonally polarized feeds p and q can be factorized into a number of physically distinct components using the Jones' matrix formalism:

$$J_i = G_i D_i P_i. \tag{6–4}$$

The first term, G_i, is usually called the "gain" and is

$$G_i = \begin{pmatrix} g_{ip} & 0 \\ 0 & g_{iq} \end{pmatrix} , \tag{6–5}$$

where g_{ip} and g_{iq} are complex gain factors for the two orthogonally polarized signals. This term represents the uncorrected effects of the atmosphere and electronics.

The second term, D_i, models imperfections in the feed polarization response. This term is given by:

$$D_i = \begin{pmatrix} 1 & d_{ip} \\ -d_{iq} & 1 \end{pmatrix} , \tag{6–6}$$

where d_{ip} and d_{iq} are complex "leakage" terms which represent the fraction of the orthogonally polarized signal "leaking" into a given feed.

The last factor in Equation 6–4 includes the effects of the rotation of an alt-az mounted antenna as seen by the source while the antenna tracks the source. This rotation, known as the parallactic angle, is given by

$$\chi = \tan^{-1}\left(\frac{\cos(\lambda)\sin(\mathrm{h})}{\sin(\lambda)\cos(\delta) - \cos(\lambda)\sin(\delta)\cos(\mathrm{h})}\right) \tag{6–7}$$

where δ is the source declination, λ is the latitude of the antenna, and h is the source hour angle. Antennas with equatorial mounts do not rotate and therefore have a constant parallactic angle (0). Parallactic angle has an effect on the measured signals which depends on the feed polarization type:
$P_i^{+} = \begin{pmatrix} \cos(\chi) & -\sin(\chi) \\ \sin(\chi) & \cos(\chi) \end{pmatrix}$ for linear or $P_i^{\odot} = \begin{pmatrix} e^{-j\chi} & 0 \\ 0 & e^{j\chi} \end{pmatrix}$ for circular feeds where $j = \sqrt{-1}$.

This assumes that the "X" (p) linear feed is oriented north–south when observing a source on the meridian and the "Y" (q) feed is rotated by 90° to the east from the "X" feed. If the feeds are rotated from this orientation, then the amount of this rotation needs to be added to χ. Since this orientation is a mechanical adjustment, its precise value may not be known and must be determined from calibration sources.

2.2. Interferometer Response

Jones/Mueller Matrices Hamaker, Bregman & Sault (1996) discuss the Mueller matrix which gives the interferometer response to polarized radiation, $v = (qq, qp, pq, qq)$. They show that the Mueller matrix is derived from the outer product of the individual antenna Jones' matrices:

$$v = (J_i \otimes J_k^*)\ S\ s \tag{6–8}$$

where s is the true Stokes visibility vector (i, q, u, v) and * indicates the complex conjugate. Matrix S is the coordinate transformation from the Stokes system to that of the correlations:
$S^{+} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & j \\ 0 & 0 & 1 & -j \\ 1 & -1 & 0 & 0 \end{pmatrix}$ for linear or $S^{\odot} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & j & 0 \\ 0 & 1 & -j & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}$ for circular.

Linearized Response In the limit of a weakly polarized source and nearly perfect feeds, any higher order terms involving source or instrumental polarization can be ignored as well as the products of such terms. The linearized approximation for crossed linearly polarized feeds on baseline $i-k$ is:

$$\begin{aligned}
v_{pp} &= \tfrac{1}{2} g_{ip} g^*_{kp} (I + Q\ \cos 2\chi + U\ \sin 2\chi),\\
v_{pq} &= \tfrac{1}{2} g_{ip} g^*_{kq} ((d_{ip} - d^*_{kq}) I - Q\ \sin 2\chi + U\ \cos 2\chi + jV),\\
v_{qp} &= \tfrac{1}{2} g_{iq} g^*_{kp} ((d^*_{kp} - d_{iq}) I - Q\ \sin 2\chi + U\ \cos 2\chi - jV),\\
v_{qq} &= \tfrac{1}{2} g_{iq} g^*_{kq} (I - Q\ \cos 2\chi - U\ \sin 2\chi), \qquad (6\text{–}9)
\end{aligned}$$

and for circularly polarized feeds:

$$\begin{aligned}
v_{pp} &= \tfrac{1}{2} g_{ip} g^*_{kp} (I + V),\\
v_{pq} &= \tfrac{1}{2} g_{ip} g^*_{kq} ((d_{ip} - d^*_{kq}) I +\ e^{-2j\chi}(Q + jU)),\\
v_{qp} &= \tfrac{1}{2} g_{iq} g^*_{kp} ((d^*_{kp} - d_{iq}) I +\ e^{2j\chi}(Q - jU)),\\
v_{qq} &= \tfrac{1}{2} g_{iq} g^*_{kq} (I - V). \qquad (6\text{–}10)
\end{aligned}$$

3. Source and Instrumental Polarization

Equations 6–9 and 6–10 show that the instrumental contribution to the cross polarized interferometer response (v_{pq} and v_{qp}) is unaffected by parallactic angle, whereas the contribution from the source does depend on the parallactic angle. In the case of circularly polarized feeds, the interferometer response is the sum of two vectors and the orientation of one varies with parallactic angle (time). If the data from interferometers using circularly polarized feeds are corrected for the phase effects of the parallactic angle, as is done for VLBI, then the source contribution is constant and the instrumental polarization rotates with parallactic angle. For interferometers with linearly polarized feeds, the situation is more complex; a point source at the phase center contributes a function of Q, U and χ to the real part of all correlations.

For interferometers with alt-az mounts (e.g. VLA, VLBA, ATCA), observations of a calibrator source over a range of parallactic angles can be used to separate the effects of source and instrumental polarization. This is illustrated in Figure 6–2a, which shows the imaginary part of the v_{qp} correlation plotted against the real part for calibrator observations made in a number of scans over a range of parallactic angle using circular feeds. The data shown in this figure have been calibrated such that the source polarization has a constant contribution whereas the instrumental polarization varies with observing geometry. The varying instrumental contribution is clearly distinguished from the source contribution. The data shown in Figure 6–2b have been corrected for the effects of instrumental polarization.

The accuracy of the instrumental polarization calibration can be limited by a number of effects, including variations of the instrumental polarization with observing geometry, instrumental phase fluctuations, ionospheric Faraday rotation and limited signal-to-noise ratio in the calibration data. Leppänen (1995) has a discussion of the effects of residual instrumental polarization errors on

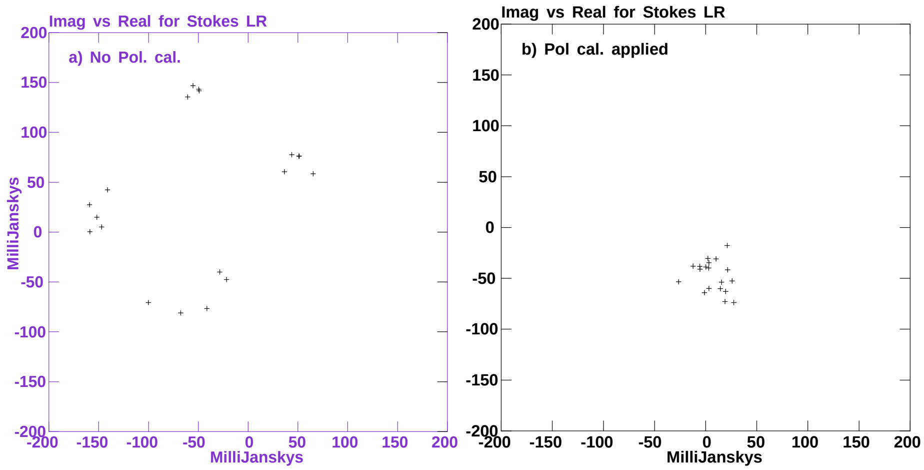


Figure 6–2. Plots of the imaginary versus real parts of the pq (left-right) correlation of the response to a weakly polarized source from a circular feed interferometer, with and without polarization calibration. The data have been rotated to remove the effects of parallactic angle, thus the source polarization is constant, and the instrumental contribution appears to rotate with time.
a) No correction for instrumental polarization, whose effects are seen to vary with parallactic angle.
b) The data shown in a) after the estimates of the instrumental polarization have been removed. Figure from Cotton (1993).

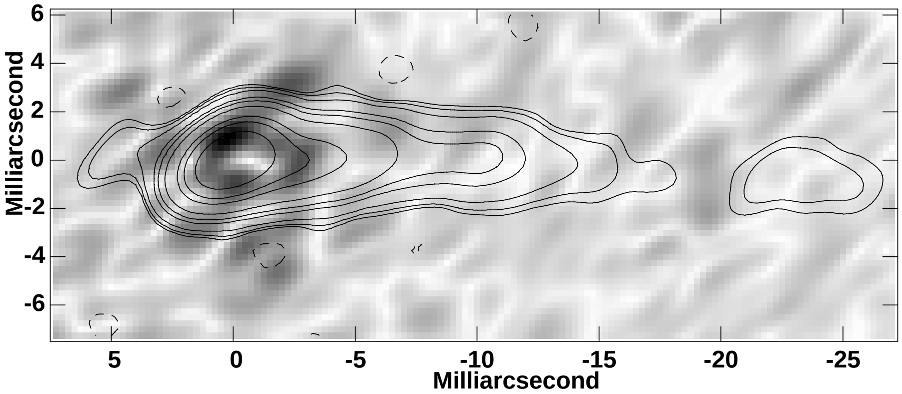


Figure 6–3. Example of an unpolarized source observed by the VLBA showing the effects of residual instrumental polarization. The contours show the total intensity of the source. The gray-scale shows the linearly polarized amplitude image after determining the instrumental polarization from the source itself. The peak polarized intensity in this image does not correspond to a feature in the total intensity (and therefore is unlikely to be real) and is about 0.5% of the peak in the total intensity.

linear polarization images. In the cases discussed by Leppänen, the source being imaged was used as its own calibrator and the result of residual instrumental polarization is the presence of weak polarized features mainly off the source. An example of this kind is shown in Figure 6–3, for the case of a strongly depolarized source. In the more general case of using a separate calibration source to determine the instrumental polarization, there will also appear to be spurious polarized emission where the total intensity is strong.

4. On-axis Calibration Strategies

This section concerns the calibration of the interferometer response at the center of the primary antenna pattern of the interferometer elements; calibration of the variations across the antenna pattern is covered in Section 5. Discussion of ionospheric Faraday rotation is deferred until Section 6.

Polarization calibration consists of several independent steps: determining and correcting instrumental polarization and calibrating the polarization angle for interferometers with circular feeds, or correcting for the mechanical alignment errors and the antenna gains errors due to calibrator polarization for antennas with linear feeds. Furthermore, polarization calibration and total intensity calibration are related and can be only partially separated.

4.1. Interaction between Polarization and Total Intensity

Astronomical calibration of the total intensity is usually performed with compact extra-galactic radio sources; either separate calibration sources or self-calibration. These sources are thought to emit via the synchrotron process and generally have a few to ten percent linear polarization and less that a half percent circular polarization. Since the best calibration sources are physically small, they tend to be variable, usually on time scales of months to years but occasionally as short as days. The net effect is that the polarization of the calibration source is usually unknown. Equations 6–9 and 6–10 show that all possible correlations are at least partially sensitive to the unknown polarization of the source, correlations of parallel circular feeds are sensitive to total intensity and circular polarization, and correlations of parallel linear feeds are sensitive to total intensity and the linear polarization. If the terms neglected in Equations 6–9 and 6–10 are considered, all correlations are sensitive to all Stokes' parameters in the detected signal; this may become important for high-dynamic-range images.

It is desirable to separate the total intensity calibration from the polarization calibration to the greatest extent possible. This is useful for both measurements involving only total intensity and to simplify the calibration of polarization sensitive data. The degree to which this is possible depends on the polarization type of the feeds involved; these issues are discussed separately in the following sections.

4.2. Circular Feeds

Correlations between parallel circular feeds are sensitive to total intensity and the source circular polarization. Synchrotron emitting, compact extra-galactic sources are quite weakly circularly polarized, usually less that 0.1%. Calibration using such a source can assume it to have no circular polarization to quite high accuracy. This approximation allows the separation of the calibration of the p and q (right and left circular) gains from each other and from the instrumental polarization.

Measured interferometer phases are sensitive only to differences in gain phases, so the usual calibration procedure is to arbitrarily set the gain phase of a "reference" antenna to zero independently for the p and q systems of feeds. Note that this independent calibration of the p and q systems of gains does not constrain the phase relationship between the two systems but ensures that the $p - q$ phase difference is that of the reference antenna. For this reason, it is important to use the same reference antenna for both the p and q calibration at a given time, and to the extent possible use the same reference antenna for all times.

The dependence of gain on parallactic angle is not explicitly shown in Equation 6–10, but a rotation of the feed rotates the phase of the response to the source. In cases such as VLBI, where the antennas have very different parallactic angles, the calibration bookkeeping is simplified if the phase effects of the parallactic angle are corrected before further calibration.

Instrumental Polarization The spurious instrumental contribution to the measured correlations are given by the "D-terms" of Equation 6–6, and instrumental polarization calibration consists of determining these values and applying corrections to the measured correlations. If the array to be calibrated has sufficiently good feeds ("D-terms" of a few percent or less), calibration source is weakly polarized (up to a few percent), and very high dynamic range is not desired, then the linearized Equation 6–10 is adequate. If these conditions are not met, then a fully nonlinear model such as Equation 6–8 is called for.

Since the polarization of the calibrator is usually unknown it must be determined jointly with the instrumental polarization. As was discussed in Section 3, parallactic angle variation on an array of alt-az mounted antennas causes a relative rotation of the complex source and instrumental contributions to the measured correlations; measurements of a source of unknown polarization over a wide range of parallactic angle is sufficient for determining both the source and instrumental polarization (see Figure 6–2).

For a long synthesis, the phase calibration source usually has a sufficient range of parallactic angle to be used in this calibration. For a shorter observation, measurements of a calibration source as it goes through transit on at least one of the interferometer elements is necessary to obtain the required range of parallactic angle.

To calibrate an array of equatorial mount antennas, a source of known (including no) polarization is required. Even for an array of alt-az mounted antennas, observations at only a single parallactic angle are required for a calibrator of known polarization. There are a few very weakly polarized compact sources which can be assumed to be unpolarized for calibration purpose.

Polarization Angle As pointed out above, the usual calibration technique of independently calibrating the p and q systems will leave an unknown phase difference between them; in the case of VLBI measurements, there will also be a delay offset between the two systems.

A $p - q$ phase difference has an equivalent effect to a change of parallactic angle, in which case Equation 6–10 shows that a constant $p - q$ phase difference will cause a rotation of $Q + jU$. As Equation 6–1 shows, this results in a rotation of the apparent polarization angle of the source. In order to calibrate the $p - q$ phase difference, or equivalently the polarization angle, sufficiently sensitive measurements of a source of known polarization angle is required. Note that a constant $p - q$ phase difference will not corrupt the derived image other than to produce the wrong polarization angle: fractional linear polarization is unaffected.

Circular polarization Equation 6–10 shows that the parallel circular correlations pp and qq respond to the total intensity plus or minus the circular polarization. The effect of incorrectly assuming that the calibration source had no circular polarization is a systematic error in the amplitude of the derived gains. Thus, to accurately calibrate the gains, a calibrator of known circular polarization is needed.

A further potential complication for circular polarization is feeds offset from the optical axis of the antenna, such as are used in the VLA and VLBA. For these antennas, there is a so-called "beam squint" in which the beams in the two polarizations are not concentric on the sky. This imposes serious limitations on the use of these arrays for circular polarization measurements; this topic comes up again in Section 5.

4.3. Linear Feeds

Correlations between parallel linearly polarized feeds are sensitive predominantly to total intensity plus a contribution from the linear polarization. The compact extra-galactic sources used for calibration purposes usually have a few to about ten percent linear polarization; in this case, the assumption of no polarization is not completely satisfactory. The effect of assuming no polarization for an unresolved calibrator and independently determining the p and q calibration is that the amplitudes of the p and q gains will be incorrect in the approximate ratio:

$$\frac{g_p}{g_q} \approx \frac{I + Q\ \cos 2\chi + U\ \sin 2\chi}{I - Q\ \cos 2\chi - U\ \sin 2\chi}, \tag{6–11}$$

which is a function of parallactic angle, hence time. If Stokes' I is formed from both the qq and pp correlations, then the effects of the erroneous assumption of no calibrator polarization cancel and the correct value of Stokes' I is obtained. Note, however, that this is not necessarily the case for self-calibration of resolved polarized sources; in this case, the data should be converted to Stokes' I before self-calibration.

The procedure described above is inadequate if polarization results are desired from the data being calibrated. A detailed description of the polarization calibration procedure used for the Australia Telescope (linear feeds and alt-az antenna mounts) is given by Sault, Killeen & Kesteven (1991).

The current discussion follows their procedure.

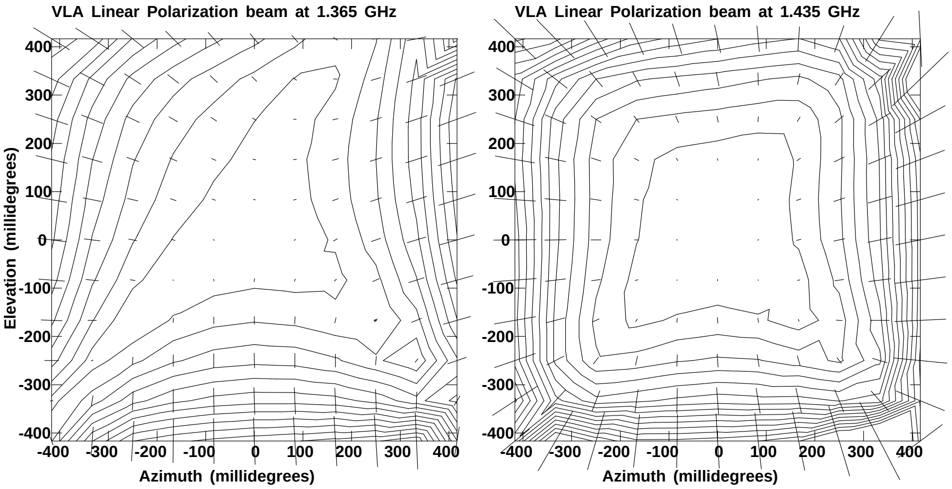


Figure 6–4. The average polarization response across the average VLA antenna pattern after correction for the on–axis instrumental polarization. a) The response at 1.365 GHz. b) The response at 1.435 GHz. Contours are shown every percent of instrumental polarization. The vectors have length proportional to the instrumental polarization and show the orientation of the E-vectors of the instrumental signals. Figure from Cotton (1994).

Calibrator and Instrumental Polarization Observations with an array using linearly polarized feeds need to include frequent measurements of an unresolved phase calibration source over a wide range of parallactic angle. If this is not possible during the normal course of the observations, then measurements of a calibrator source at a declination near the latitude of the array as it transits the meridian may suffice. The variable response to the calibrator's linear polarization allows separation of the source and instrumental contributions to the correlations. Sault, Killeen & Kesteven (1991) describe an iterative technique of alternately solving for antenna gains and then the source and instrumental polarization.

Feed Orientation/$p-q$ phase Measurements of a weakly polarized source, especially when its polarization is unknown, do not allow the determination of any errors in the assumed orientation of the feeds or of the $p-q$ phase difference. The determination of these parameters requires observations of a strongly polarized source of known polarization, such as 3C 138 or 3C 286.

5. Wide Field Polarization

The preceding discussion only relates only to the instrumental response at the center of the array element primary antenna pattern. In the general case, the instrumental polarization varies across the primary antenna pattern. Figure 6–4 shows this effect for the "average" VLA antenna at two nearby frequencies. In this case, there is significant off-axis instrumental polarization in regions of the antenna pattern which have sufficient gain to be useful for imaging extended

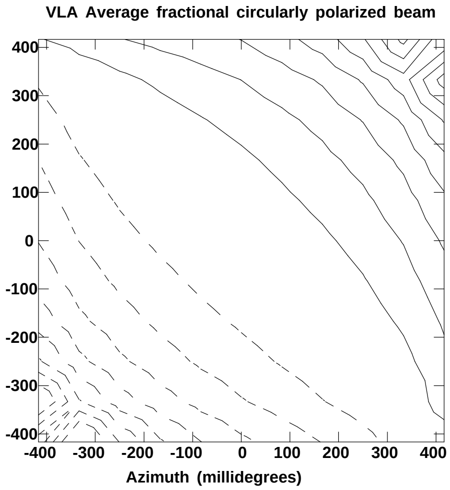


Figure 6–5. The spurious circularly polarized response of the average VLA antenna across its primary beam. Contours are shown every 10 percent of spurious instrumental circular polarization; negative contours are shown dashed. Figure from Cotton (1994).

sources or large fields of view. Figure 6–4 also shows that these effects are strong functions of observing frequency.

The pattern of off-axis polarization is fixed to the antenna and rotates on the sky with parallactic angle. This rotation of the instrumental polarization pattern on the sky will tend to average out its effects on the image derived from a long synthesis; however, any correction must be applied to the visibility data rather than to the resultant image. For snapshot images where there is not a significant variation of parallactic angle, or when using equatorial mounts, a correction for off-axis instrumental linear polarization can be applied to the final image. See Cotton (1994) for a detailed discussion of this technique as it applies to the VLA.

The spurious instrumental contribution to circular polarization may also vary across the antenna pattern, especially for off-axis circularly polarized feeds. An example of this which graphically illustrates the difficulties of circular polarization measurements with the VLA is shown in Figure 6–5.

6. Ionospheric Faraday Rotation

As a linearly polarized electromagnetic wave propagates through an magnetized plasma, its orientation rotates — an effect called Faraday rotation. An alternate way of describing this effect is that the speed of propagation of the right and left circularly polarized components of the radiation are different from each other in a magnetized plasma. The different velocities cause the relative phases of the right and left circular polarizations to change.

Faraday rotation of emission from the region around or in front of the source may be of interest to the astronomer, but Faraday rotation in the Earth's ionosphere seldom is. In addition, the ionospheric Faraday rotation varies with the ionization and recombination of the ionosphere and as the path through the Earth's magnetic field changes with observing geometry. The effect of Faraday rotation on the observed polarization angle is given by (Pacholczyk 1970):

$$\Delta\Phi \;=\; \frac{0.93 \;\times\; 10^6 \int N \; H_{||} \; ds}{(2\pi\nu)^2} radians \qquad (6\text{–}12)$$

where N is the electron density (cm^{-3}), $H_{||}$ is the component of magnetic field parallel to the direction of propagation (gauss), ν is the radio frequency (Hz) and s is distance along the line of sight (cm).

Since ionospheric Faraday rotation is time variable, it can corrupt polarization images from an extended synthesis or produce incorrect polarization angles for snapshot observations. The strong dependence on observing frequency means that it is most troublesome for low frequency observations, especially near maxima in solar activity when the ionosphere is the most active. There are rapid variations of the ionosphere near sunrise and sunset, or when the ejecta from a solar flare hits the earth. Any observations at frequencies below about 2 GHz may potentially be affected.

6.1. External Calibration

Corrections for ionospheric Faraday rotation may be determined from measurements of the total electron content (TEC) in the earth's atmosphere made using satellites (GPS these days) or estimated from the mean sunspot number (Chiu 1975). Using an assumed profile of the ionosphere and a model of the Earth's magnetic field, it is possible to estimate the Faraday rotation in a given direction.

Measurements of Faraday rotation toward cosmic sources may also be used to determine the TEC. However, since the Faraday rotation depends on the path through the ionosphere to the source, the observed Faraday rotation towards a calibrator cannot be directly applied to a target source unless the two are very close on the sky (the maximum separation depends on the amount of Faraday rotation).

Ionospheric Faraday corrections used for the VLA are derived from a dual frequency GPS receiver. The delay differences between the 1575 and 1227 MHz signals are used to determine the TEC along the line of sight to each satellite. These measurements are then fitted to a model of the ionosphere, which gives the zenith TEC at the location of the GPS receiver and gradients in latitude and longitude. This technique is discussed in Erickson et al. (1999).

6.2. Self Calibration

A variation on self calibration may be applied to determining the time varying relative Faraday rotation in the direction of the source, provided the source has sufficient polarized emission. A "reference" polarization image can be derived from data collected during a sufficiently short period that the ionospheric Faraday rotation is essentially constant. The visibility data can then be divided into time segments over which the Faraday rotation can be assumed to be constant; the Faraday rotation in each segment is determined by comparison with the "reference" polarization image. If there are strongly polarized, isolated features

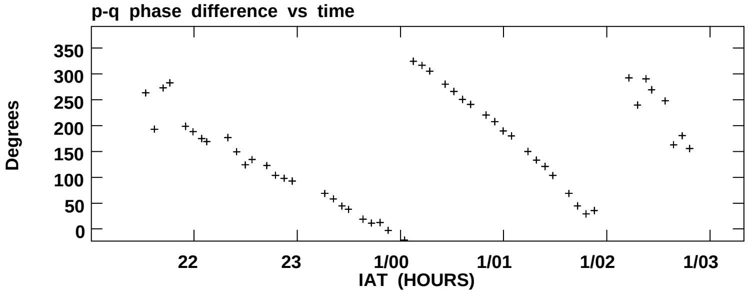


Figure 6–6. The $p - q$ phase difference derived for VLA observations of a source at 333 MHz. Several turns of phase variations are seen over the period of the observations. Figure from W. Cotton (1995)

in the source, this comparison may be done using images made from each time segment. If the polarized emission is more extended, then it may be necessary to determine the Faraday rotation using the Fourier transform of the "reference" polarization image and the visibility data in each time segment. Note that this procedure will remove the variable effects of the Faraday rotation but will not correct the polarization angles derived from the data. Fractional polarization and relative polarization angle will be preserved, but the absolute polarization angle will not. This is analogous to the loss of position information in the use of self calibration to calibrate the gain phases.

An example of the results of this procedure is shown in Figure 6–6 which shows the derived $p - q$ (R-L) phase difference as a function of time. These observations were made at a relatively low frequency and the Faraday rotation induced more than two full turns of $p - q$ phase during the course of the measurements. If uncorrected this would have very seriously corrupted the derived polarization image.

7. Imaging

Imaging of polarization data is quite similar to that of total intensity data, although there are a few differences that merit discussion. Once the calibration parameters are known, equation 6–8 or 6–9 or 6–10 can be inverted to derive the I, Q, U, and V visibilities, which can then be imaged and deconvolved using your favorite technique. One major difference between I and the polarized Stokes parameters is that I is always positive, whereas the other Stokes parameters may have negative values. This means that positivity may not be used as a constraint on the polarization deconvolution.

Conversion of the measured correlations to the Stokes' parameters generally requires all four (pp, qq, pq, and qp) correlations. There is a special case of observations using circularly polarized feeds in which some observations have only one of the pq or qp correlations. In this case, $Q + jU$ can be imaged and Q and U images recovered using a complex deconvolution. See Cotton (1993) for a more detailed discussion of this technique.

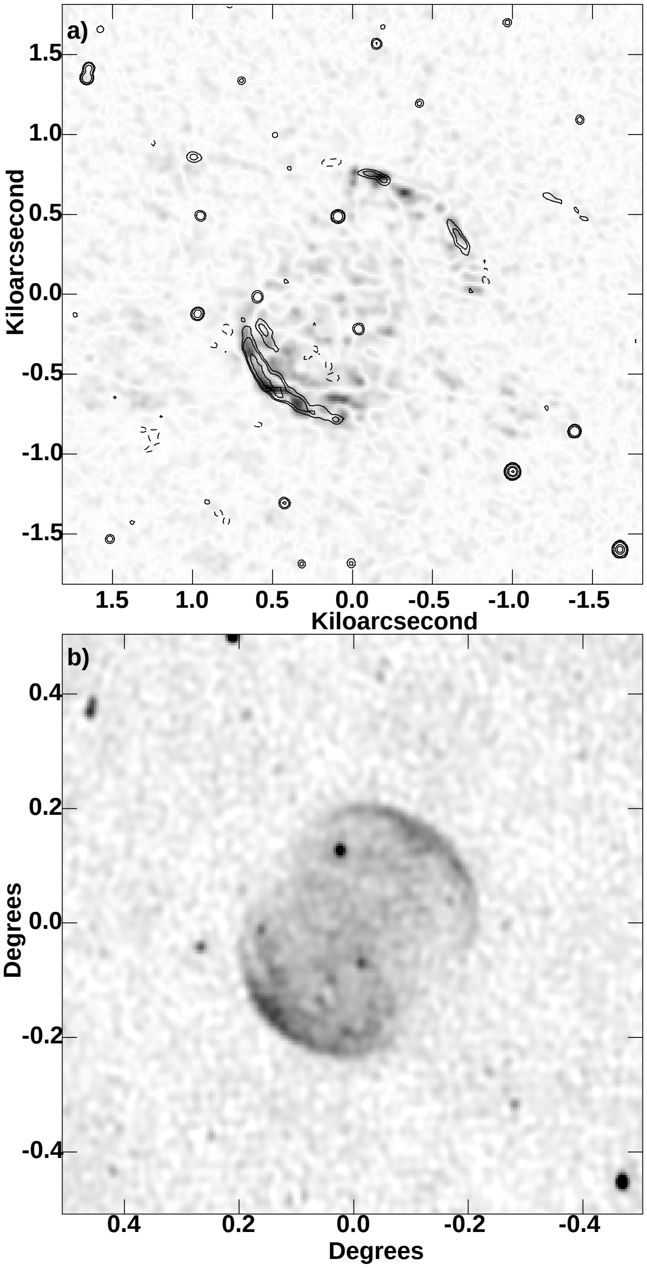


Figure 6–7. An example of the different spatial frequency filtering effects of interferometers on total intensity and linear polarization.
a) Total intensity contours superimposed on a gray scale representation of the linearly polarized emission. Polarized emission is visible in regions where total intensity is not. These images are from the NVSS survey (Condon et al. 1998) which does not well sample the spatial frequencies in this object.
b) The total intensity image of the same region of sky shown in a) but from the WENSS survey (Rengelink et al. 1997) which has approximately the same resolution but better sampling of the lower spatial frequencies. In this figure, there is visible total intensity emission in the regions of polarized emission of a).

In the true emission from a source, the total intensity cannot be less than the sum of all polarized components. This is not necessarily the case in images derived from interferometer arrays in which the sky has been subjected to a spatial frequency filtering. Structure on size scales that are strongly resolved by all interferometer baselines tend to be lost in the images derived from these measurements. The total intensity may be smooth on a size scale that is resolved by all interferometers, whereas the polarized emission may not be if there are variations in fractional polarization or polarization angle — perhaps due to variable Faraday rotation in the source — across this region. In this case, it is possible to have images which appear to exceed 100% polarization or to have polarized emission in regions where no total intensity is visible.

An example of this effect is shown in Figure 6–7a, in which an extended source appears to have polarized emission where there is no total intensity. The total intensity image of the same region of the sky at similar resolution but made using better short baseline coverage is shown in Figure 6–7b. In Figure 6–7b, there is visible total intensity emission everywhere that polarized emission appears in Figure 6–7a.

Acknowledgments. I would like to thank Fred Schwab for many helpful discussions, and Chris Flatters for information about Faraday rotation corrections at the VLA.

References

Born, M. & Wolf, E. 1975, *Principles of Optics*, Pergamon Press, Oxford

Chandrasekhar, S. 1960, *Radiative Transfer*, Dover Press, New York.

Chiu, Y. T. 1975, *J. At. Terr. Phys.*, 37, 1573.

Condon, J. J., Cotton, W. D., Greisen, E. W., Yin, Q. F., Perley, R. A., Taylor, G. B., & Broderick, J. J. 1998, *AJ*, 115, 1693.

Conway, R. G. & Kronberg, P. P. 1969, *MNRAS*, 142, 11.

Cotton, W. D. 1993, *AJ*, 106, 1241.

Cotton, W. D. 1994, AIPS Memoranrum No. 86.

Cotton, W. D. 1995, in *Very Long Baseline Interferometry and the VLBA* eds. J. A. Zensus, P. J. Diamond & P. J. Napier (San Francisco: PASP), 289–306.

Erickson, W., Perley, R. A., Flatters, C., Kassim, N., & Payne, J. 1998, submitted to *A&A*.

Fomalont, E. B. & Wright. M. C. H. 1974, in *Galactic and Extra–Galactic Radio Astronomy* eds. G. L. Verschuur, & K. I. Kellermann (Berlin: Springer Verlag), 256.

Hamaker, J. P., Bregman, J. D., & Sault, R. J. 1996, *A&AS*, 117, 137.

Kraus, J. D. 1966, *Radio Astronomy*, McGraw-Hill, New York.

Leppänen, K. 1995, PhD Thesis, Helsinki University of Technology.

Morris, D., Radhakrishnan, V., & Seielstad, G. A. 1964, *ApJ*, 139, 551.

Rengelink, R. B. et al. 1997, *A&AS*, 124, 259–280.

Sault, R. J., Hamaker, J. P., Bregman, J. D. 1996, *A&AS*, 117, 149.

Sault, R. J., Killeen, N. E. B., & Kesteven, M. J. 1991. “AT Polarization Calibration”, AT Technical Document Series 39.3/015.

Schwab, F. 1979, VLA Computer Memorandum No. 154.

Stokes, G. 1852, *Trans. Cambridge Phil. Soc.*, 9, part 3, 399–416.

Thompson, A. R., Moran, J. M., & Swenson, G. W., Jr. 1986, *Interferometry and Synthesis in Radio Astronomy.* New York: Wiley–Interscience. First (1991) and second (1994) reprintings by Krieger Pub. Co., Malabar (Florida).

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

7. Imaging

Daniel S. Briggs
NCSA, Champaign-Urbana, IL 61801, U.S.A.

Frederic R. Schwab
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A

Richard A. Sramek
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. This lecture covers formation of the estimated sky brightness via linear methods. The formalism of the dirty image is developed from the fundamental Fourier transform relationship between observed visibility and sky brightness. The practical computational approximation to this formalism is then covered in detail. Several weighting schemes used to control the shape of the dirty beam are presented. The convolutional gridding used to interpolate the irregularly sampled data onto a rectangular grid is examined in detail, including aliasing of sources outside the primary field of view and ramifications of the choice of convolutional gridding function.

1. Fourier Transform Imaging

A fundamental result of Lectures 1 and 2 was the existence of a Fourier transform (FT) relationship between the sky brightness I, the primary beam pattern $\mathcal{A}$, and the visibility V observed with an interferometer. From Lecture 2 (Eq. 2–26),

$$\mathcal{A}(l,m)I(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{7–1}$$

This simple relation holds if (a) $\left|\frac{\Delta\nu}{c}\,\mathbf{b}\cdot(\mathbf{s}-\mathbf{s_0})\right| \ll 1$ and (b) $|w(l^2+m^2)| \ll 1$. These conditions are met whenever the radiation to which the interferometer pairs respond originates in a suitably small (and confined) region of sky. Since the correction for the primary beam can be made trivially at the final stage of data processing[1] (as discussed in Lecture 1, Sec. 4.4), we shall use $I(l,m)$ to denote the *modified sky brightness*, $\mathcal{A}(l,m)I(l,m)$.

V is complex-valued and, after the usual calibration steps (see Lecture 5), is reckoned in units of flux density ('Janskys', 1 Jy $= 10^{-23}$ ergs cm^{-2} s^{-1} Hz^{-1}), while I has units of surface brightness (flux density per unit of solid angle). A standard unit for I is Jy/beam area; sometimes Jy per square arcsecond is used instead. The units are determined by the normalization of Equation 7–1.

Equation 7–1 is used to obtain an estimate of the modified sky brightness from the observed visibilities, recorded at (u,v) points (u_k,v_k), $k=1,\ldots,M$. In practice, M may range from ten to a few hundred with a two element interferometer, to over a million with a multi-element array like the VLA. With M

[1] This is assuming that $\mathcal{A}$ has been carefully measured over a large enough region in (l,m). Wide-field imaging, in cases in which a source covers, say, a larger region than the central lobe of the primary beam, is an especial problem. Antennas with azimuth–elevation mounts (as at the VLA) also present a problem because the primary beam patterns rotate on the sky, as functions of parallactic angle. See Lecture 6.

modest, model fitting is feasible—and sometimes useful (see Lecture 16). But for large M the usual method of estimating I is via the discrete Fourier transform (the DFT), because extremely efficient algorithms are known for numerical evaluation of DFTs.

The topics of some of the lectures to follow also fall under the broad category of 'imaging'. But the discussion here is restricted to 'simple-minded' methods of estimating the sky brightness: that is, *directly* approximating the right-hand side of Equation 7–1, via only *linear* operations. The so-called 'dirty image' that results is a discrete approximation to I^D, where (from Lecture 1, Eq. 1–10)

$$I^D(l,m) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u,v)V'(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{7–2}$$

Here, S denotes the (u,v) sampling function and V' the observed visibility; the prime indicates that the visibility data are noise-corrupted measurements. (For conciseness, I^D has been left unprimed, but it too is noise-corrupted whenever V is.)

1.1. The 'direct Fourier transform' and the FFT

Either of two methods is commonly used to numerically approximate the Fourier transform in Equation 7–2. The first, called the 'direct Fourier transform' method,[2] approximates $I^D(l,m)$ by brute-force evaluation of the sum

$$\frac{1}{M}\sum_{k=1}^{M} V'(u_k,v_k)e^{2\pi i(u_k l+v_k m)}\,. \tag{7–3}$$

If this 'direct Fourier transform' is evaluated at every point of an $N \times N$ grid, the number of real multiplications required is $4MN^2$ (the number is halved, though, assuming Hermitian data). In practice M is usually of the same order as N^2, so the number of multiplications goes roughly as N^4. The number of sine and cosine evaluations required is also $\mathcal{O}(N^4)$, as is the number of additions/ subtractions.

The second method requires interpolating the data onto a rectangular grid, so that a fast Fourier transform (FFT) algorithm can be used. The process of interpolation is referred to as *gridding*. (Gridding may require sorting the

[2] This choice of terminology is unfortunate. The natural abbreviation for the term—'DFT'—is used almost universally (by everyone except radio astronomers) to stand for something else: the 'discrete Fourier transform'. For example, the 2–D discrete FT of an $M \times N$ matrix (x_{ij}) is the $M \times N$ matrix (y_{kl}) given by

$$y_{kl} = \sum_{p=1}^{M}\left(e^{2\pi i(p-1)(k-1)/M}\sum_{q=1}^{N} x_{pq}e^{2\pi i(q-1)(l-1)/N}\right)\,.$$

The major distinction between the two usages is that in one case the data are regularly spaced, and in the other they are not. Also, the 'direct FT' is generally not invertible, whereas the 'discrete FT' is; usually the term 'transform' is reserved for invertible transformations.

data into order of decreasing $|u|$ or decreasing $|v|$.) The number of elementary arithmetic operations required by the technique most often used for gridding is $\mathcal{O}(M)$. The number of such operations required by an FFT algorithm (say, the Cooley–Tukey algorithm) is only a few times $N^2 \log_2 N$—not $\mathcal{O}(N^4)$! This saves much computing time for large databases, and large N especially, if an economical method of interpolation is used. However, for making small images (i.e., for N small) from small databases (M small), the 'direct Fourier transform' may be faster than the combination of gridding and FFT.

In the following sections we first discuss weighting and selection of (u, v) data and how it affects the resulting images. This applies no matter how the Fourier transform is approximated. Then we touch upon the problems that are introduced by gridding the data to permit use of the FFT—the problems of aliasing and correction for gridding.

2. The Sampling Function, and Weighting the Visibility Data

The sampling function S and its Fourier transform, the synthesized beam B, were introduced in Lecture 1. In practice the data are variously weighted, according to their reliability and to control the shape of the synthesized beam.

2.1. The sampling function

S is a 'generalized function', or 'distribution', which may be expressed in terms of the two-dimensional Dirac delta function, or 'δ-distribution',

$$S(u,v) = \sum_{k=1}^{M} \delta(u - u_k, v - v_k) \,. \tag{7–4}$$

It is useful to introduce a second generalized function, called the *sampled visibility function* or, alternatively, the (u, v) *measurement distribution*,[3]

$$V^S(u,v) \equiv \sum_{k=1}^{M} \delta(u - u_k, v - v_k) V'(u_k, v_k) \,. \tag{7–5}$$

That is, $V^S = SV'$. Let $\mathfrak{F}$ denote the Fourier transform operator. Equation 7–2 can be rewritten

$$I^D = \mathfrak{F} V^S = \mathfrak{F}(SV') \,. \tag{7–6}$$

By the *convolution theorem*, which says that the Fourier transform of a product of functions is the convolution of their FTs (see, e.g., Bracewell 1978),

$$I^D = \mathfrak{F} S * \mathfrak{F} V' \,, \tag{7–7}$$

[3] Note that the visibility measurements are not, in actuality, point samples of the inverse Fourier transform of the modified sky brightness $\mathcal{A}I$, but that instead they represent *local averages* of it. Time- and frequency-averaging, which are discussed in Lecture 2, are the dominant averaging effects. One should try to choose observing parameters (integration time and bandwidth) that make relatively safe our assumption here about δ-function sampling. This matter is further discussed in Lectures 17 and 18.

where $*$ denotes convolution. For a point source of unit strength, centered at position (l_0, m_0), $|V'(u,v)| \equiv 1$ (plus noise), and $\mathfrak{F}V'$ is the (shifted) Dirac δ-function: $\mathfrak{F}V'(l,m) = \delta(l - l_0, m - m_0)$. So the point source response of the array, i.e., the *synthesized beam*, is given by $B = \mathfrak{F}S * \delta = \mathfrak{F}S$. Equation 7–7 is the familiar result (Lecture 1, Eq. 1–11) that the observed brightness is the true brightness convolved with this 'beam'.

It should be apparent that the so-called 'direct Fourier transform', as defined by Equation 7–3, is *exactly* I^D. That is to say, that—assuming δ-function sampling—$I^D(l,m)$, as defined by Equation 7–2, is given exactly by the discrete summation Eq. 7–3. Equation 7–7 holds exactly for the 'direct Fourier transform' method, (an analogous relation is given below for the FFT method). Of course, a computed 'direct Fourier transform' image is indeed an approximation, but only in the sense that it is inevitably a discretely sampled version of I^D and that the sums are computed in finite precision arithmetic.

2.2. Weighting functions for control of the beam shape

In analogy to Equation 7–4, a *weighted sampling function*, or *weighted sampling distribution*, can be written as

$$W(u,v) = \sum_{k=1}^{M} R_k T_k D_k \delta(u - u_k, v - v_k)\,. \tag{7–8}$$

And, in analogy to Equation 7–5, one can define a *weighted, sampled visibility function*, or *weighted and sampled measurement distribution*, V^W according to $V^W = WV'$, or, explicitly,

$$V^W(u,v) = \sum_{k=1}^{M} R_k T_k D_k \delta(u - u_k, v - v_k) V'(u_k, v_k)\,. \tag{7–9}$$

The coefficients R_k, T_k, and D_k (discussed below) are weights assigned the visibility points. These data points may represent time-averages of visibility measurements spaced along the loci of the (u,v) tracks. R_k is a weight that indicates the reliability of the k^{th} visibility datum. It may depend on the integration time, the system temperature, and the bandwidth used for that data point.

There is no control over the value of R_k in the image formation, and one might hope to ignore it here. Yet the manner in which the data samples are combined will influence the sensitivity of the final map as discussed in Lecture 8. It is an unfortunate reality that the data weighting which produces the most desirable beam from an imaging standpoint will often utilize the data very irregularly and result in poor sensitivity—the subjects of imaging and sensitivity are inextricably linked. The procedure for best balancing the desirable properties of low & uniform sidelobes, high resolution, and high sensitivity for a given project is complicated and still somewhat heuristic, though some progress has been made towards reasonable compromises that work well in the majority of cases. Here we present several examples showing the effect of different parameters on the weighting, but be aware that the best strategy for a given project may use more than one in combination. See Briggs (1995) for a more exhaustive treatment of different weighting strategies.

The full data calibration may not be available at the imaging stage, but it is often the case that the thermal variation of the data sample is the dominant contribution to R_k or that the non-random components of R_k can be ignored. In this case, it can be assumed that R_k is proportional to the inverse variance of the sample distributions of $\operatorname{Re} V_k'$ and $\operatorname{Im} V_k'$. The factor by which the point source sensitivity of the output dirty map is degraded by the choice of the T_k and D_k can then be calculated. This has been called the weighting noise or the normalized thermal RMS, and is given by

$$WT_{\text{noise}} = \Delta I^D / \Delta I^D_{\text{best}} = \sqrt{\Big(\sum_{k=1}^{M} T_k^2 D_k^2 R_k\Big)\Big(\sum_{j=1}^{M} R_j\Big)} \Big/ \sum_{i=1}^{M} T_i D_i R_i \qquad (7\text{–}10)$$

See Lecture 8 or Briggs (1995) for the full derivation. Modern imaging programs now often display this quantity, giving the careful user quantitative feedback about the effect of the weighting on sensitivity.

If S were a smooth, well-behaved function—say, a Gaussian—then B would have no sidelobes, just smooth 'wings'. In practice, S is a linear combination of many δ-functions, often with gaps in the (u, v) coverage corresponding to missing interferometer spacings. There is always a finite limit to the extent of the (u, v) coverage, corresponding to the largest (projected) spacing of interferometer elements. In addition, for most arrays more data points fall in the inner region of the (u, v) plane than further out. This tends to give higher weight to the low spatial frequencies. Thus the natural sampling may impair effective deconvolution or mask interesting features of I. All of these real world concerns combine to produce a beam which is rarely what the astronomer wishes. The density weight D_k and the taper T_k are completely arbitrary and can be specified in many Fourier transform imaging programs, to 'fine-tune' the beam shape and combat the natural sampling as best possible. They are factored into two independent functions purely for convenience in specification. The T_k are used to downweight the data at the outer edge of the (u, v) coverage, and thus to suppress small-scale sidelobes and increase the beamwidth. The D_k are used to offset the high concentration of (u, v) tracks near the center, and to lessen the sidelobes caused by gaps in the coverage; i.e., to simulate more uniform (u, v) coverage. We shall discuss these forms of weighting separately.

The tapering function The T_k are specified by a smooth function T: $T_k = T(u_k, v_k)$. T is usually separable, so that $T(u, v) = T_1(u)T_2(v)$; and often it is a radial function (i.e., a function with circular symmetry): $T_k = T(r_k)$ where $r_k \equiv \sqrt{u_k^2 + v_k^2}$. Although functions whose radial profiles follow a power-law or powers of a cosine are occasionally used, the most prevalent form is the Gaussian. The dispersion, or the half-width at half amplitude, or the half-width at 0.30 amplitude are used in different data reduction programs to specify the characteristic width (or widths) of T (see Fig. 7–1). The modern trend has been towards specification of the taper by the equivalent convolution with a Gaussian function in the image plane, so Fig. 7–1 also gives a short table of conversions between the two conventions.

For a Gaussian taper, $T(r) = \exp(-r^2/2\sigma^2)$, the half-power beamwidth (i.e., the width of the synthesized beam, measured between half-amplitude points)

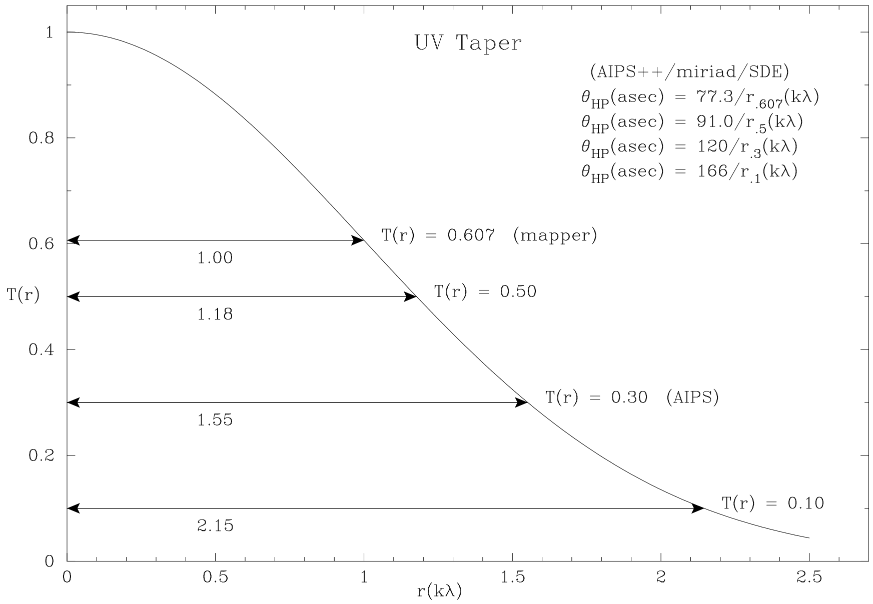


Figure 7–1. A Gaussian (u, v) taper with dispersion $\sigma = 1$ km.

is $\theta_{\rm HPBW} = 0.37/\sigma$ with θ in radians and σ in wavelengths. Translated into common units, $\theta_{\rm HPBW} = 0.77\lambda_{\rm (cm)}/\sigma_{\rm (km)}$ arcseconds. This holds only for a densely sampled Gaussian that is not truncated by the edge of the (u, v) coverage. When the taper is negligible at the edge of the (u, v) coverage (assuming dense coverage), one can use a filled circular aperture approximation, for which $\theta_{\rm HPBW} = 2.0\lambda_{\rm (cm)}/a_{\rm (km)}$ arcseconds, where a is the radius of the aperture. Real-life observational geometries and (u, v) coverages often produce larger $\theta_{\rm HPBW}$ and, frequently, elongated beams. Examples of the VLA point source response with different (u, v) tapers are shown in Figure 7–2.

Instead of de-emphasizing data near the outer boundary of the (u, v) coverage, it is sometimes desirable to downweight the data near $u = v = 0$. An undersampled large-scale emission region may introduce large undulations in image intensity that are hard to remove. These can present a problem for detecting a weak point source embedded within a region containing extended emission. Minimum (u, v) limits and other forms of downweighting are often used to diminish the effect of these low spatial frequency data points.

Finally, while one normally thinks of tapering as downweighting the visibility data as a function of radius, it is also possible to inverse taper and upweight the higher spatial frequencies instead. An upweighting has no equivalent finite convolution in the image plane, but it can arise in the solution of a convolution equation between two Gaussians. If one wishes to form a Gaussian beam of a given shape—say for matching resolution between images of an object at two epochs—one can solve the equation $B_{\rm target} = \mathit{taper} * B$ for the equivalent taper in the visibility plane. This might involve an inverse taper, but can still yield reasonable results if the upweighting is not too severe. Unfortunately, few imaging programs are able to do this yet, but the capability will likely become more

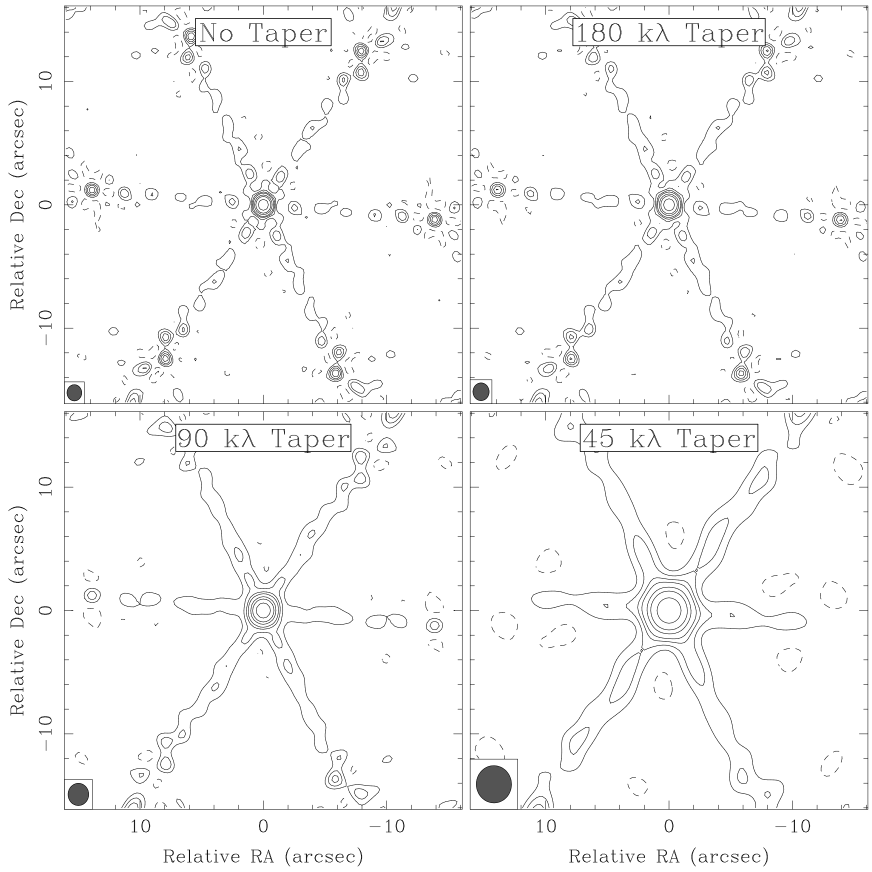


Figure 7–2. The effect of a Gaussian taper on the point source response of a VLA snapshot in the **A** configuration at 20-cm wavelength. As a narrower Gaussian taper (i.e., a heavier tapering) is applied, the half-power width of the point spread function increases and the inner sidelobes are reduced.

common in the future. Details of this formalism are given in Appendices B–D of Briggs 95.

The density weighting function: The density weighting function can be used to compensate for the clumping of data in the (u, v) plane by weighting by the reciprocal of the local data density. Two choices for this weighting are commonly provided:

$$D_k = 1\,, \qquad \text{called } \textit{natural weighting}, \qquad (7\text{–}11)$$

$$\text{and} \quad D_k = \frac{1}{N_s(k)}\,, \qquad \text{called } \textit{uniform weighting}, \qquad (7\text{–}12)$$

where $N_s(k)$ is the number of data points within a symmetric region of the (u, v) plane, of characteristic width s, centered on the k^{th} data point. (s might be the radius of a circle or the width of a square.)

Natural weighting, with all points treated alike, gives the best signal-to-noise ratio for detecting weak sources. However, since the (u, v) tracks tend to spend more time per unit area near the (u, v) origin, natural weighting emphasizes the data from the short spacings, and tends to produce a beam with a broad, low-level plateau. This latter feature is especially undesirable when imaging sources with both large-scale and small-scale structure.

With uniform weighting, a common choice for N_s is to count all the points that lie within a rectangular block of grid cells in the neighborhood of the k^{th} datum (gridding is discussed later).[4] This produces a beam specified largely by the tapering function T. In the case where different points have different reliability weights R_k, the $N_s(k)$ is usually replaced by the total reliability weightsum in that same region. In most Fourier transform imaging programs s is a free parameter selected by the user. The default value of s is usually a function of the physical image dimension, so by merely changing the image or pixel size, one is also changing the uniform density weights.

Sometimes, especially in the VLA 'snapshot' mode of observing, uniform weighting may not be 'uniform' enough. Although all cells have equal weight, the filled cells are still concentrated toward the center and along the arms of the VLA 'Y'. At the further expense of signal-to-noise ratio, the size parameter s can be increased. This 'super-uniform weighting' gives lightly sampled, isolated cells weights comparable to those given cells in well-sampled parts of the plane. The result is again a beam shape controlled more by the tapering function and less by the arrangement of the sampled visibilities. Examples of the VLA point source response obtained with various weighting functions are shown in Figure 7–3.

A hybrid form of the uniform and natural weighting called *robust weighting* has recently been introduced, which arises from a minimization of the summed sidelobe power and thermal noise. A typical tradeoff between beam resolution and weighting noise for a full track VLA observation, traced by varying the robustness parameter, is shown in Figure 7–4. Characteristically this tradeoff curve is an 'L' shape, meaning that one can profitably work either in the knee of the curve for a compromise beam with intermediate properties in both parameters, or work on one leg of the 'L' and slightly improve one parameter without greatly affecting the other. This weighting scheme is currently available in most packages and is discussed extensively in Chapter 3 of Briggs (1995). As with the other weighting parameters described, the most appropriate robustness parameter for a given dataset must be determined empirically.

3. Gridding the Visibility Data

To take advantage of the extreme efficiency of the FFT algorithm, visibility values must be assigned to a regular, rectangular matrix or 'grid', usually with a power-of-two number of points along each side. Since the observed data seldom lie on such a grid, some procedure (an interpolation procedure comes most readily to mind) must be used to assign visibility values at the grid points, based on

[4] In the AIPS implementation, these blocks are called 'uniform weight boxes', and the size of the weight box determines the degree of super-uniform weighting. In AIPS++/miriad/SDE, super-uniform weighting is specified with a Field of View (FOV) parameter. In most respects, an AIPS box size of N is equivalent to an FOV of 1/N.

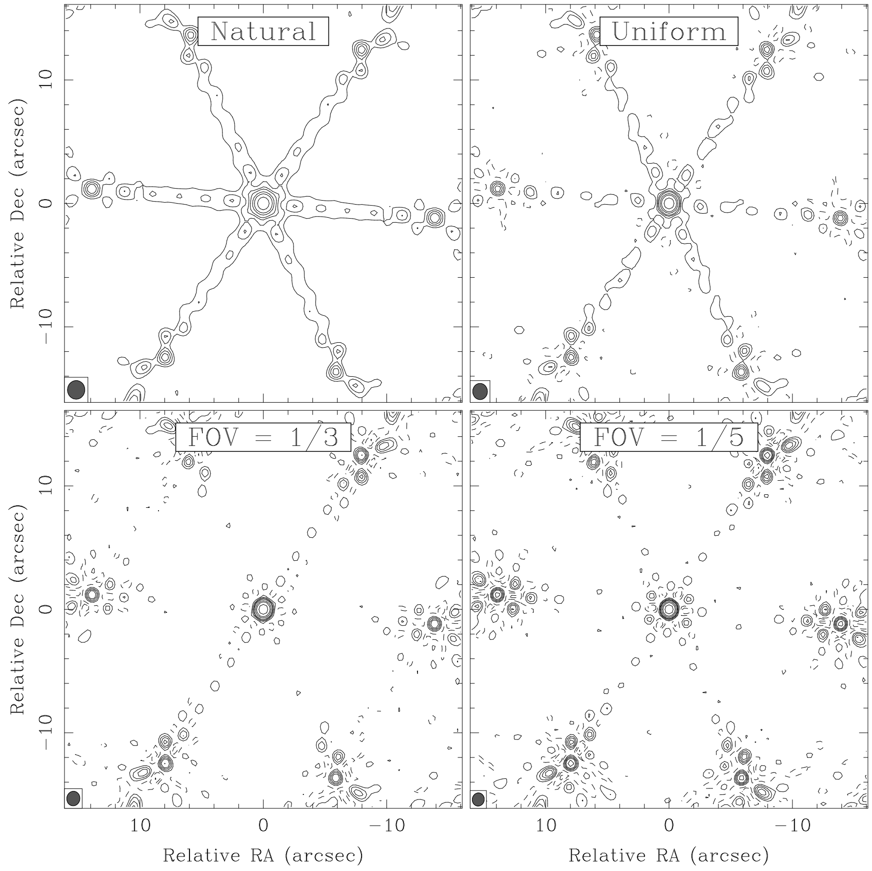


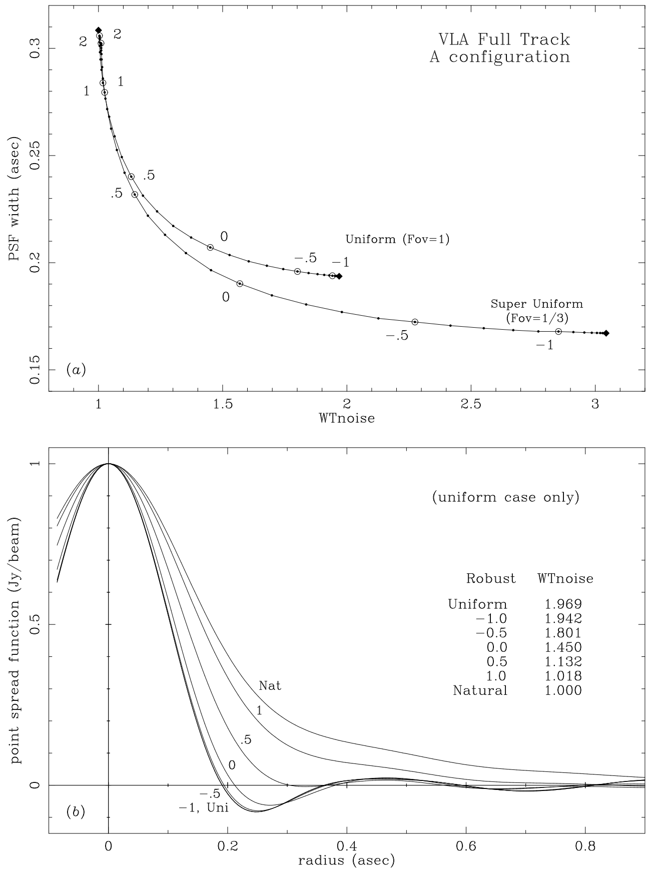
Figure 7–3. The effect of different weighting functions on a VLA 'snapshot' image of a point source.

the observed values.[5] There are many ways to achieve this interpolation (see, e.g., Thompson & Bracewell 1974), but with quasi-randomly placed observations a convolutional procedure in the (u, v) plane leads to an image with predictable distortions and to results that are easy to visualize. Convolution is not, in fact, a pure interpolation procedure, since it combines smoothing, or averaging, with interpolation. This should not be viewed as undesirable—given that there often are many noisy, possibly discrepant, data points in the neighborhood of a given grid point.

3.1. Gridding by convolution

The idea is to convolve the weighted, sampled measurement distribution V^W with some suitably chosen function C, and to sample this convolution at the

[5] Some special array geometries (e.g., 'T's and Crosses, with elements aligned linearly N–S and E–W) can provide regularly spaced data. See, for example, the description of the Clark Lake array by Erickson *et al.* (1982). The assumption (mentioned below) of a sufficiently large number of data points in the neighborhood of each filled 'cell' is not required. However aliasing problems persist, because of the regular sampling.



Robust	WTnoise
Uniform	1.969
−1.0	1.942
−0.5	1.801
0.0	1.450
0.5	1.132
1.0	1.018
Natural	1.000

Figure 7–4. The effect of robustness parameter on weighting noise and beam shape for a full track VLA observation. The Uniform and Robust=−1 traces in panel **(b)**, are visually identical, with the Robust=−.5 trace just above them. Notice that all plotted traces are distinct in WTnoise, however.

center of each 'cell' of the grid. For economy's sake—and because it seems reasonable for the value assigned at a given grid point to equal some local average of the measurements— C, in practice, is always taken to be identically zero outside some small, bounded region A_C. Since V^W is a linear combination of M δ-functions, this convolution $C * V^W$, evaluated at the grid point (u_c, v_c), is given by

$$\sum_{k=1}^{M} C(u_c - u_k, v_c - v_k) V^W(u_k, v_k)\,. \tag{7–13}$$

Note that, since the region A_C is quite small in area, there are generally *many* fewer than M nonzero terms in this sum.

Note also that Eq. 7–13 does not, in fact, represent a *local average* of the measurements in the neighborhood of (u_c, v_c). For that, some sort of normalization would be required—say, multiplication by the area of A_C, followed by division by the number of data points whose shifted coordinates $(u_c - u_k, v_c - v_k)$ lie within the region A_C (and one would want C to integrate to unity). When this particular form of normalization is used, the normalized sum (ignoring weighting) approaches the *non-discrete, integral convolution* $C * V$ evaluated at (u_c, v_c) as the number of measurements increases without bound, provided that the measurements in the neighborhood of (u_c, v_c) are uniformly distributed, and provided that the noise in V' is well-behaved. In practice, this straightforward form of normalization is not always incorporated in imaging—so the matter of normalization becomes intertwined with that of 'density weighting', discussed above.

The operation of sampling $C * V^W$ at all points of the grid may be represented by the equation

$$V^R = R\left(C * V^W\right) = R\left(C * (WV')\right), \tag{7–14}$$

where (as usual) multiplication is indicated by juxtaposition and where R, a 'bed of nails' resampling function, is given in terms of Bracewell's 'sha' function (denoted Ш) by

$$R(u,v) = \text{Ш}(u/\Delta u, v/\Delta v) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \delta(j - u/\Delta u, k - v/\Delta v)\,. \tag{7–15}$$

Here, Δu and Δv define the cell size—i.e., the separation between grid points. This operation is called *resampling* (hence the R-notation) because, as you recall, the interferometer array earlier provided the samples embodied in V^S and V^W. Now, since V^R is a linear combination of regularly spaced δ-functions, a matrix of samples of its Fourier transform $\mathfrak{F}V^R$ can be obtained by a discrete Fourier transform. Thus $\mathfrak{F}V^R$ can be calculated by the FFT algorithm.

$\mathfrak{F}V^R$—after normalization, and after one simple correction—is what you have been seeking: a 'dirty' image—a cheap approximation to I^D. Denote $\mathfrak{F}V^R$ by $\widetilde{I}^D$.

Applying the convolution theorem to Equation 7–14, $\widetilde{I}^D$ is given by

$$\widetilde{I}^D = \mathfrak{F}R * \left[(\mathfrak{F}C)\left(\mathfrak{F}V^W\right)\right] = \mathfrak{F}R * \left[(\mathfrak{F}C)\left(\mathfrak{F}W * \mathfrak{F}V'\right)\right]\,. \tag{7–16}$$

(Please refer now to Fig. 7–5 for a graphical interpretation of Eq. 7–16 and for an illustration of the operations that are described in the remainder of this section.) Ш is its own Fourier transform; R behaves similarly—by the dilation property of the FT (see Sec. 4.1.),

$$(\mathfrak{F}R)(l,m) = \Delta u\, \Delta v \text{Ш}(l\Delta u, m\Delta v) = \Delta u\, \Delta v \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \delta(j - l\Delta u, k - m\Delta v)\,. \tag{7–17}$$

One effect of the resampling is to make $\widetilde{I}^D$ a periodic function of l and m, of period $1/\Delta u$ in l and period $1/\Delta v$ in m. Another effect, called *aliasing*, is also introduced. It, too, arises because of the convolution with the scaled sha function $\mathfrak{F}R$ (more on this later, in Sec. 3.2.).

The FFT algorithm generates one period of (a discrete version of) $\widetilde{I}^D$. To image a rectangular region of width $N_l \Delta\theta_l$ radians in l and $N_m \Delta\theta_m$ in m, one chooses grid spacings satisfying $N_l \Delta u = 1/\Delta\theta_l$ and $N_m \Delta v = 1/\Delta\theta_m$ wavelengths. An $N_m \times N_l$ FFT yields the discretely sampled version of $\widetilde{I}^D$. Let P denote the region over which $\widetilde{I}^D$ is computed— i.e., P, which is called the *primary field of view*, is given by $|l| < N_l \Delta\theta_l/2$, $|m| < N_m \Delta\theta_m/2$.

The net effect of the gridding convolution is to multiply the sky brightness by a function $c(l,m)$, the FT of the convolving function C (i.e., $c \equiv \mathfrak{F}C$). The tapering function T, introduced earlier for control of the beam shape, has the effect of a convolution in the image domain.

An image representing the point source response of the array, or the 'dirty beam' B^D, can be obtained by setting all the measurements $V'(u_k, v_k)$ to unity and following the steps outlined above. Denote the image so obtained by $\widetilde{B}^D$.

Normally, $\widetilde{I}^D$ and $\widetilde{B}^D$ are corrected for the effect of the gridding convolution by pointwise division by c: The so-called 'grid-corrected' image is given by

$$\widetilde{I}^D_c(l,m) = \frac{\mathfrak{F}R * \left[(\mathfrak{F}C)\,(\mathfrak{F}V^W)\right]}{\mathfrak{F}C} = \frac{\widetilde{I}^D(l,m)}{c(l,m)}\,, \tag{7–18}$$

and the 'grid-corrected' beam by

$$\widetilde{B}^D_c(l,m) = \frac{\widetilde{B}^D(l,m)}{c(l,m)}\,. \tag{7–19}$$

The commonly used term 'grid corrected' is, in a way, a misnomer, since one is actually correcting for the effect of the convolution function C. The grid correction is not an exact correction, except in the limit of a large number of well-distributed visibility measurements. It also is not exact due to the presence of R in Equation 7–14 and $\mathfrak{F}R$ in Equation 7–16. It could be so only if $c(l,m)$ were identically zero outside of the region being imaged; this is impossible because C is confined to a bounded region A_C.[6]

Finally, $\widetilde{I}^D_c$ and $\widetilde{B}^D_c$ both are normalized by a scaling factor selected so that the peak of $\widetilde{B}^D_c$ is of unit flux density. One may as well not alter the notation to reflect this, since it is a trivial operation.

If $c(l,m)$ tends sufficiently rapidly to zero outside P, so that the resampling can be ignored, and if the (u,v) samples are well enough distributed for the gridding correction to be approximately valid, then $\widetilde{I}^D_c$ is a good approximation to I^D— that is, Equation 7–16 becomes

$$\widetilde{I}^D_c = \mathfrak{F}W * \mathfrak{F}V'\,, \tag{7–20}$$

—and then the usual convolution relation between I^D, B, and I is approximately valid with $\widetilde{I}^D_c$ and $\widetilde{B}^D_c$ substituted for I^D and B, respectively. Note, however,

[6] The FT of any nontrivial (i.e., nonzero) function which is confined to a bounded region has features extending to infinity. By a theorem of Paley and Wiener (see, e.g., Dym & McKean 1972) the FT of such a function is extremely well-behaved, in the sense that it can be analytically extended to an entire function in the complex domain (i.e., in the case of 2 dimensions, from $\mathbf{R}^2$ to $\mathbf{C}^2$). In particular, the FT cannot vanish over any open set (this is why the synthesized beam has sidelobes that 'never go away').

that $\widetilde{B}_c^D$ is usually computed only over a region of the same dimensions as the image $\widetilde{I}_c^D$. For this reason, the deconvolution algorithms (described in Lecture 8) usually operate just on a region with one-quarter the area of the input image.

Figure 7–5 *(pp. 140–141).* A graphical illustration of the steps in the imaging process is shown in this one-dimensional example. At the top, in panels **(a)** and **(b)**, a model source and its visibility are displayed side-by-side; the results of successive imaging operations are displayed vertically. The image domain is shown on the left, and the visibility domain on the right. Horizontally opposed panels represent Fourier transform pairs. The units on the vertical axes were chosen arbitrarily— i.e., we have not bothered with normalization. The horizontal axes are in radians for the image domain plots, at left; the baselines are expressed in wavelengths for the visibility domain plots, at right.

The model source, shown in panel **(a)**, is the sum of a Gaussian-shaped extended source and four symmetrically placed point sources. The total flux density of the Gaussian is 1.5 times the sum of the fluxes in the point sources. This symmetry was chosen to ensure that the visibility function, shown in panel **(b)**, is real-valued and even, allowing a simpler display. Panel **(d)** shows the telescope transfer function, or sampling function S, which includes a central 'hole'. We have chosen a smooth function for simplicity, but one should note that no array would in fact produce a smooth sampling function. In reality, S is a sea of closely- and irregularly-spaced δ-functions, as in Equation 7–8. The triangular sampling density was chosen to mimic the fall-off in the density of samples with increasing spacing. The telescope beam B corresponding to **(d)** is shown in panel **(c)**. The data available for imaging are shown in panel **(f)**; this product of the true visibility function and the sampling function corresponds to V^S, as defined by Equation 7–9. The image which a direct transformation of **(f)** would yield is shown in panel **(e)**. This image is equal to the convolution of the beam **(c)** with the true sky brightness **(a)**. This image shows a large amplitude oscillation, reaching a negative peak centered on the position of the extended source. This effect, which is of much larger amplitude than the oscillation seen in **(c)**, is due to the missing central spacings in the (u, v) sampling and to the fact that the visibility of an extended source is relatively highly concentrated near $u = v = 0$. With sufficient computing resources (mammoth resources would often be required), one might use the 'direct Fourier transform' method of Section 1.1.; **(e)** is the image that would result.

Extra steps are required to make use of the FFT: First, the data are convolved with some suitably chosen function, and then they are resampled over a regularly-spaced grid (in practice the convolution is evaluated only at the grid points). For illustration, a simple, and crude, convolution function C was employed, as shown in **(h)**. The sharp drop-off in C creates large, oscillating wings in its Fourier transform, shown in **(g)** (the image-plane representation of the 'grid-correction function'). The data, after convolution, are shown in panel **(j)**. If a (continuous) Fourier transform were applied at this stage, the result would appear as in panel **(i)**. The important effect to note is that the outermost point sources have been inverted in amplitude. This occurs because the convolution function that we have chosen is too wide. The inner point sources have been slightly reduced in amplitude, though not inverted in sign. As the FFT requires regularly spaced data, the data in **(j)** must be sampled. The (re-)sampling function R is shown in panel **(l)**, and its transform, the replication function, in panel **(k)**. The resampled, convolved visibility is shown in panel **(n)**. These are the data that the FFT actually sees. The FT of this is the image shown in panel **(m)**; it has been replicated at the various points shown in panel **(k)**. Notice that aliases of the outermost point sources appear just outside the positions of the innermost point sources. This aliasing occurs because the resampling function, shown in panel **(l)**, undersamples (i.e., takes fewer than 2 samples per cycle) of the transform of the outermost point sources. The final operation is correcting for the effect of the convolution. This is done by dividing the image by the Fourier transform of the convolution function. For this example, only the region of the image where the inner lobe of $c > .1$ has been retained, though this is not an issue in practice. The result is shown in panel **(o)**. This is the end product, the 'dirty image' that is supplied to the deconvolution programs.

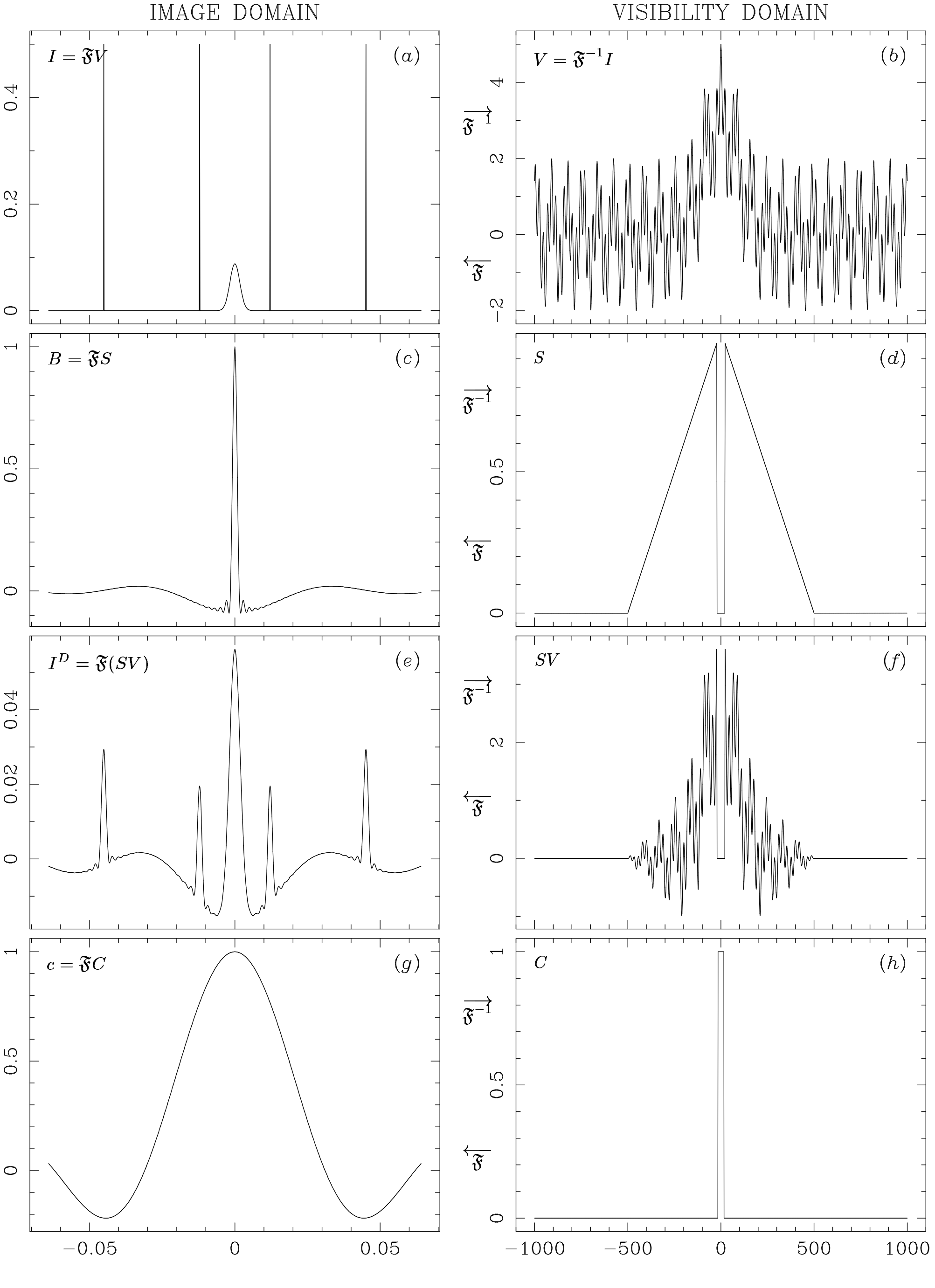


Figure 7–5. *(Caption is on p. 139).* (Continued on next page).

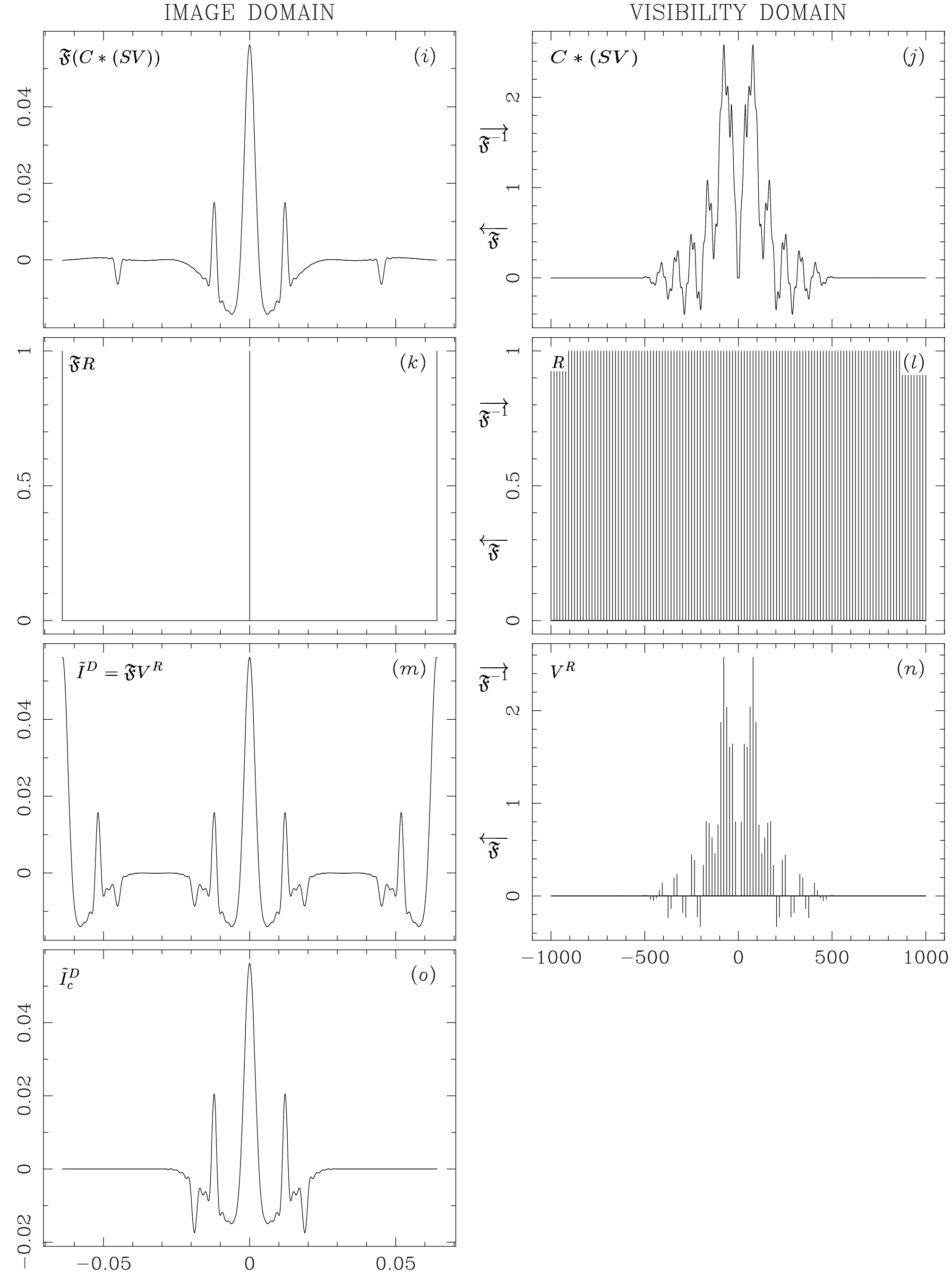


Figure 7–5. *(Continued).*

3.2. Aliasing

Due to the presence of $\mathfrak{F}R$ in Equation 7–16 and to the fact that c is not identically zero outside the primary field of view, parts of the sky brightness that lie outside P are aliased, or 'folded back', into P. Undersampling, and the truncation of the sampling at the boundaries of the (u, v) coverage, are the root causes of aliasing. (If the sky brightness I has features extending over a region of width Ω_l in l and width Ω_m in m, then its visibility function has been undersampled if the visibility samples are separated by more than $1/\Omega_l$ in u and $1/\Omega_m$ in v.) The amplitude of an aliased response from position (l, m) is determined by $|c(l, m)|$. The simplest way to tell whether a feature is aliased or authentic is to calculate images with different cell sizes $\Delta\theta$; an aliased feature then appears to move, while a real one stays the same angular distance from the image center. Additionally, an image covering the full main lobe of the primary beam may quickly reveal whether there is an aliasing problem in an image of a smaller region.

Aliasing of sources that lie outside the primary field of view is only part of the problem. Although it may be possible to obtain visibility samples that are closely enough spaced to avoid undersampling over the sampled region of the (u, v) plane, the finite physical size of the array sets a limit on how far the sampling can extend. For this reason, any authentic feature within P has sidelobes extending outside the image. These sidelobes are also aliased into P, effectively raising the background variance and resulting in a beam shape that depends on position. If, for example, the visibility function is well sampled over a square region of the (u, v) plane but no samples are obtained outside that region, then (assuming uniform weighting) the sidelobes in I^D are precisely those of Gibbs' phenomenon, discussed in Lecture 4.

3.3. Choice of a gridding convolution function

The best ways to avoid aliasing problems are (a) to make the image large enough that there are no sources of interest near the edges of the image, (b) to avoid undersampling, and (c) to use a gridding convolution function C whose Fourier transform c drops off very rapidly beyond the edge of the image. Desideratum (c) favors gridding convolution functions that are not highly confined in the (u, v) plane. But, in practice, computing time restricts one's choice of C to functions that vanish outside a small region, typically six or eight (u, v) grid cells across. A compromise must be struck between alias rejection and computing time. Most programs to date have used a width of six cells, though the modern trend may be moving towards eight.

C is always taken to be real and even. And, since C is usually separable—i.e., $C(u, v) = C_1(u)C_2(v)$ —we shall continue the discussion in just one dimension. Typical choices for C are:

- a 'pillbox' function,
- a truncated exponential,
- a truncated sinc function ($\operatorname{sinc} x \equiv \frac{\sin \pi x}{\pi x}$),
- an exponential multiplied by a truncated sinc function, and

- a truncated spheroidal function.

Each is truncated to an interval of width m grid cells, so that $C(u) \equiv 0$ for $|u| > m\Delta u/2$; thus $\mathcal{O}(Mm^2)$ arithmetic operations are required for gridding. These functions are described below; for more discussion see Schwab (1978, 1980):

- *Pillbox.* $C(u) = \begin{cases} 1\,, & |u| < m\Delta u/2\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$ For $m = 1$, convolution with this C is equivalent to simply summing the data in each cell. Calculation of these sums is fast, but the alias rejection is the worst of the five functions considered here. c is a scaled sinc function.

- *Exponential.* $C(u) = \exp\left[-\left(\frac{|u|}{w\Delta u}\right)^{\alpha}\right]$. Typically $m = 6$, $w = 1$, and $\alpha = 2$. That is, a truncated Gaussian is often used, in which case c can be expressed in terms of the error function.

- *Sinc.* $C(u) = \operatorname{sinc}\frac{u}{w\Delta u}$. Typically $m = 6$, $w = 1$. c can be expressed in terms of the sine integral. If m is allowed to increase, c approaches a step function that is constant over P and zero outside. This is the intuitive justification for considering the use of this function, that the FT of a unit step function truncated at $\pm\frac{1}{2}$ is the sinc function.

- *Exponential times sinc.* $C(u) = \exp\left[-\left(\frac{|u|}{w_1\Delta u}\right)^{\alpha}\right] \operatorname{sinc}\frac{u}{w_2\Delta u}$. Typically[7] $m = 6$, $w_1 = 2.52$, $w_2 = 1.55$, $\alpha = 2$; i.e., a truncated, Gaussian-tapered sinc function is often used. c can easily be computed by numerical quadrature, but it lacks a closed-form expression.

- *Spheroidal functions.* $C(u) = |1 - \eta^2(u)|^{\alpha}\psi_{\alpha 0}(\pi m/2, \eta(u))$, with $\psi_{\alpha 0}$ a 0-order spheroidal function (Stratton 1935), $\eta(u) = 2u/m\Delta u$, and $\alpha > -1$. For $\alpha = 0$ this is the 0-order 'prolate spheroidal wave function', which is the optimal C (among all square-integrable functions of width m grid cells) in that the energy concentration ratio $\int_P |c(l)|^2\, dl \Big/ \int_{-\infty}^{\infty} |c(l)|^2\, dl$ is maximized. The other $\psi_{\alpha 0}$ are optimal in the sense of maximizing a *weighted* concentration ratio: for given α, $\int_P w(l)|c(l)|^2\, dl \Big/ \int_{-\infty}^{\infty} w(l)|c(l)|^2\, dl$ is maximized, where $w(l) = |1 - 2l\Delta u|^{\alpha}$. Choosing $\alpha > 0$ gives higher alias rejection near the center of the image, at the expense of alias rejection near the edges. ψ_{00} is its own FT, in the sense that if you truncate it as done here, and then take the FT, what you get back is ψ_{00}. Similarly, the other $\psi_{\alpha 0}$ are finite Fourier self-transforms, in the sense that if you so truncate one, weight it, and transform it, what you get back is $\psi_{\alpha 0}$. $\psi_{\alpha 0}$ is used at the VLA, with $m = 6$ and $\alpha = 1$ being typical. See Schwab (1984) for further discussion and additional references.

[7] For a gridding convolution function of this particular parametric form, these values of the characteristic widths w_1 and w_2 are an optimal choice, in the sense described below in the discussion of ψ_{00}.

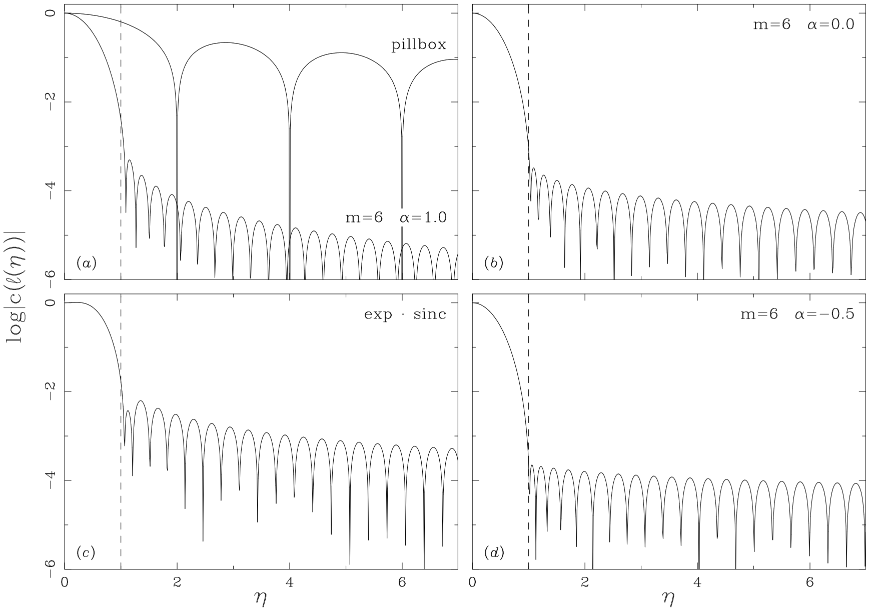


Figure 7–6. For some typical gridding convolution functions C, plots of the absolute value of the Fourier transform of C. **(a)** The spheroidal function ψ_{10}, for $m = 6$, compared with the pillbox function ($m = 1$); **(b)** the 'prolate spheroidal wave function' ψ_{00}, $m = 6$; **(c)** an optimized Gaussian-tapered sinc function, $m = 6$; **(d)** the spheroidal function $\psi_{-\frac{1}{2},0}$, $m = 6$. Panel **(a)** is comparing the function most commonly used at the VLA with the simple but particularly poor choice of a pillbox. Adapted from Schwab (1984).

Figure 7–6 shows the Fourier transforms of various typical gridding convolution functions, normalized to unity at $l = 0$. The abscissa on this plot is in units of image half-widths, $\eta = 2l\Delta u$, so that $\eta = \pm 1$ at the image edges. The image response is suppressed at the edge for both functions, however the exp · sinc function is flatter inside P, and drops much faster past the image edge. The aliased response can, of course, be negative, producing an apparent 'hole' in the image.

The plots in Figure 7–6 compare the pillbox function and the Gaussian-tapered sinc function with several spheroidal functions. The quantity of most direct importance is the ratio of the intensity of an aliased response to the intensity the feature would have if it actually lay within the primary field of view P, at the position of its alias: if η' denotes the position within P at which the aliased response of a source at position η appears, then this suppression ratio is given by $q(\eta) = |c(l(\eta))/c(l(\eta'))|$. (And η' is given by $\eta' = ((\eta + 1) \bmod 2) - 1$; it is useful to sketch a plot to convince oneself of this.) The suppression ratio for the same functions as in Figure 7–7 is given in Figure 7–7.

The pillbox, exponential, and sinc functions do not give as effective alias rejection as the exp · sinc or the spheroidal. The exp · sinc has somewhat smaller corrections and, thus smaller errors (due to round-off noise and to violation of the assumptions that make the grid correction valid), near the image edges, while the spheroidal has better rejection beyond the image edge (Schwab 1984).

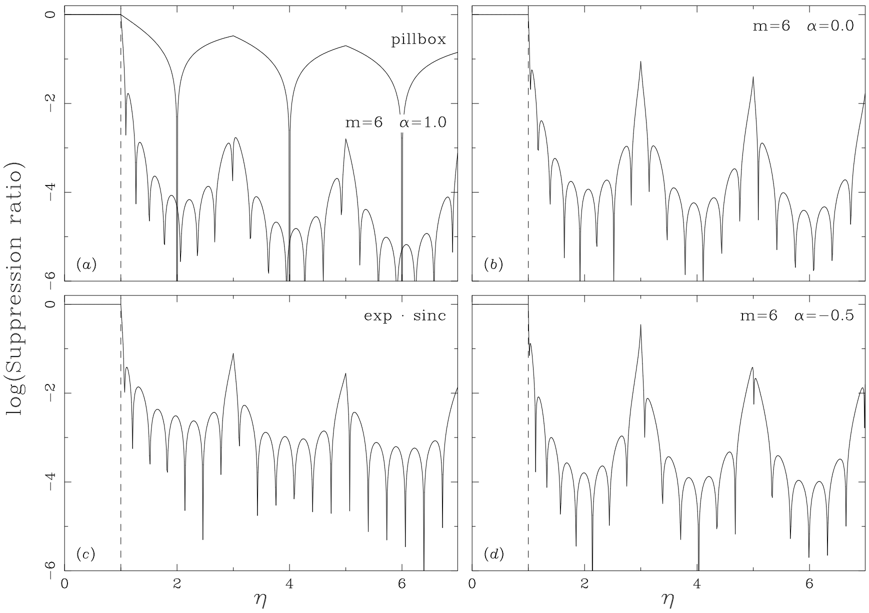


Figure 7–7. The suppression ratio for the same convolution functions as in Figure 7–6.

Remember that the convolution functions suppress only aliased responses. Sidelobes which legitimately fall within the primary field of view, whether from sources inside or outside P, are not suppressed (see Fig. 7–8). With alias suppression of 10^2 to 10^3, at two or three image half-widths, it is these sidelobes which may cause the dominant spurious image features and impair effective deconvolution.

4. Additional Topics

4.1. Translating, rotating, and stretching images

The Fourier transform possesses three basic symmetry properties that are useful in radio interferometric imaging. The first important property is the behavior of the Fourier transform with respect to translation—that is, with respect to a shift of origin: namely, if you shift a function, i.e., replace $f(\mathbf{u})$ by $f(\mathbf{u} - \Delta\mathbf{u})$, and take the FT you get the same result as if you had first taken the FT and then multiplied by $e^{2\pi i \mathbf{x}\cdot\Delta\mathbf{u}}$ (here $\mathbf{x}$ denotes the variable in the transform domain). Similarly, if you want a shift of origin $\Delta\mathbf{x}$ in the transform domain, all you need do is multiply, before transforming, by a factor $e^{-2\pi i \mathbf{u}\cdot\Delta\mathbf{x}}$. Thus, in imaging, all that is required to achieve a shift of origin in the image is to multiply the visibilities by the appropriate complex exponentials before transforming.

The second important property is that the Fourier transform commutes with rotations; that is, if you take the FT and then rotate the coordinate system in the transform domain, you get the same result as if you had first rotated the coordinate system and then taken the FT. Thus, to 'turn an image around', all that you need do is rotate the (u, v) coordinates of the visibility data. (It is easy

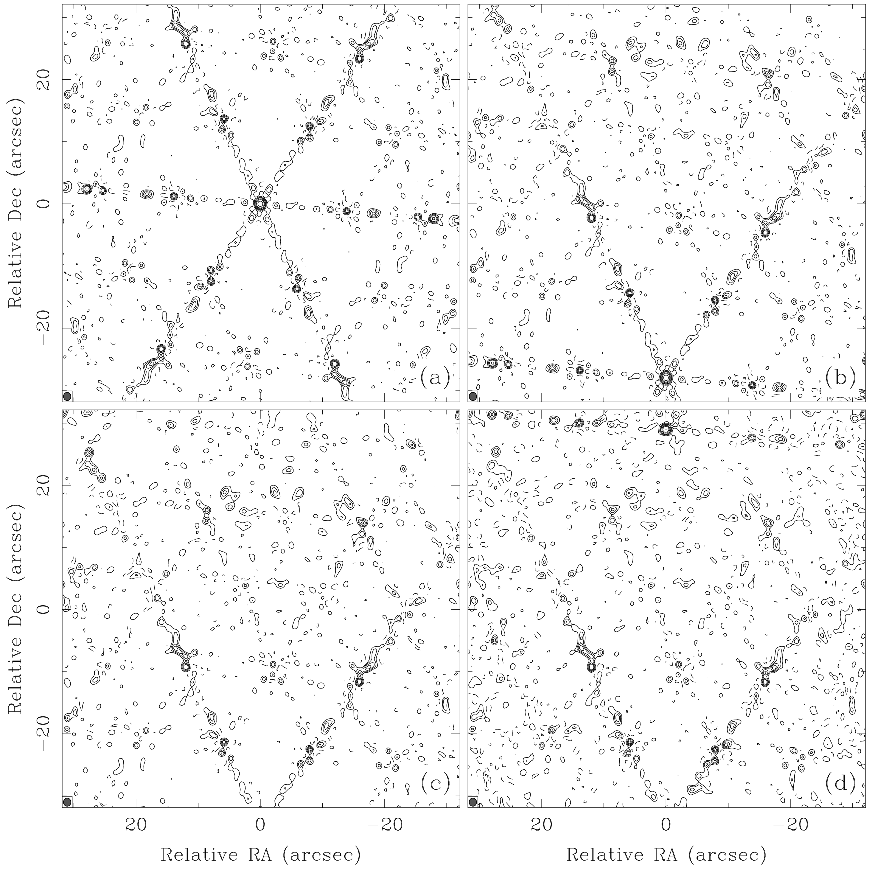


Figure 7–8. The effects of aliasing: **(a)** a point source at the field center using the standard spheroidal convolution; **(b)** the same source near the image edge; **(c)** the same source below the image edge—the sidelobe response is unchanged, but there is no obvious aliased response to the source; **(d)** the source below the lower image edge, but using the pillbox convolution function—a dramatic aliased image of the source appears at the top edge.

to see why the FT has this property: the inner product $\mathbf{u} \cdot \mathbf{x}$ in the exponential kernel of the FT is invariant under rotation.) At the VLA, the visibility (u, v) coordinates are routinely rotated to correct the data for differential precession—i.e., to put the data into the coordinate reference frame of a standard epoch, say, J1950 or J2000. Data taken at two different epochs, say a year apart, need this correction for differential precession before they can be sensibly combined or compared; routine correction to a standard epoch automatically rectifies this problem. Additionally, it is sometimes convenient to rotate the coordinate system so that features in a source have a particular alignment in an image. For an elongated source, this can reduce the data storage requirements (by reducing the number of pixels needed to represent the source by a computed, discrete image) and therefore aid during deconvolution (see Lecture 8) by reducing the required number of arithmetic operations.

The third basic symmetry property of the FT is that it *anti*-commutes with dilations. That is, if you 'stretch' a function linearly and isotropically, then its FT 'shrinks' proportionately. (That is, the FT of $g(\mathbf{u}) = f(\alpha\mathbf{u})$ is given by $(\mathfrak{F}g)(\mathbf{x}) = \alpha^{-n}(\mathfrak{F}f)(\mathbf{x}/\alpha)$. The multiplicative constant α^{-n} depends on the dimensionality n.) Or, if you linearly stretch a function in just one coordinate, then its FT 'shrinks' proportionately, but in only one of the coordinate directions. This property is the reason that, for a fixed array geometry, the spatial resolution increases (i.e., the characteristic width of the synthesized beam beam decreases) with observing frequency — as the (u, v) coverage expands, the beam shrinks proportionately.

Following Bracewell (1978), the shift property is sometimes called the *shift theorem*, and the dilation property the *similarity theorem.*

4.2. Practical details of implementation

Many Fourier transform imaging programs do not work quite as described above. Sometimes the tapering, introduced in Equation 7–8, and specified by $T(u, v)$, is applied after gridding. This would appear to make only a minute difference. But, in the same sense in which it is incorrect to ignore resampling to justify the grid correction, it is also incorrect to ignore the convolution with $\mathfrak{F}T$, which, if inserted into Equation 7–16, would now appear outside the square brackets.

For economy, Fourier transform imaging programs often do not attempt to evaluate the gridding convolution function very accurately, but instead use a step function (tabular) approximation, with steps spaced at increments of, typically, $\Delta u/100$. This introduces another (not very serious) 'replication' effect like that due to $\mathfrak{F}R$, but one with a very long period, $100/\Delta u$. The grid correction given by Equation 7–18 should be based now on the FT of the step function approximation to C rather than on the FT of C itself. For analysis, see Greisen (1979). (Schwab (1984) gives cheap and accurate rational approximations to the spheroidal functions; the step function approximation is unnecessary.)

4.3. Non-coplanar baselines

In Equation 7–1 the visibility samples are expressed as a function of two variables, u and v, rather than as a function of (u, v, w). As shown in Section 6 of Lecture 2, Equation 7–1 is strictly valid whenever the visibility measurements are confined to a plane, as they would be if obtained with an interferometer array whose elements are aligned along an East–West line; and, again as shown in Lecture 2, this relation is approximately valid when $I(l, m)$ is confined to a small region of sky—that is, when our condition (b) holds, $|w(l^2 + m^2)| \ll 1$. In wide-field imaging with non-coplanar baselines, condition (b) is often violated.

Recall from Lecture 2 (Eq. 2–21) the relation

$$V(u,v,w) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \frac{\mathcal{A}(l,m)I(l,m)}{\sqrt{1-l^2-m^2}} e^{-2\pi i\left(ul+vm+w\left(\sqrt{1-l^2-m^2}-1\right)\right)}\, dl\, dm\,. \quad (7\text{–}21)$$

This can be rewritten as

$$V(u,v,w)e^{-2\pi i w} = \quad (7\text{–}22)$$

$$\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \frac{\mathcal{A}(l,m)I(l,m)}{\sqrt{1-l^2-m^2}}\, \delta(n-\sqrt{1-l^2-m^2}) e^{-2\pi i(ul+vm+wn)}\, dl\, dm\, dn\,.$$

Now, by sampling V, weighting by $e^{-2\pi i w}$ and the Fourier kernel, and integrating over (u,v,w), one obtains an analog of Equation 7–2,

$$I^{D\,(3)}(l,m,n) = \tag{7–23}$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u,v,w)V(u,v,w)e^{-2\pi i w}e^{2\pi i(ul+vm+wn)}\,du\,dv\,dw\,.$$

This is equal to a familiar looking three-dimensional convolution:

$$I^{D\,(3)} = I^{(3)} * B^{D\,(3)} \tag{7–24}$$

with

$$I^{(3)}(l,m,n) \equiv \frac{\mathcal{A}(l,m)I(l,m)}{\sqrt{1-l^2-m^2}}\,\delta(n-\sqrt{1-l^2-m^2})\,, \tag{7–25}$$

and

$$B^{D\,(3)}(l,m,n) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u,v,w)e^{2\pi i(ul+vm+wn)}\,du\,dv\,dw\,. \tag{7–26}$$

Note that $I^{(3)}$ is a distribution confined to the celestial sphere embedded in a three dimensional volume and that $B^{D\,(3)}$ is mostly concentrated near the origin, i.e., near $l=m=n=0$.

Either of the methods described earlier for approximating I^D can be extended straightforwardly to Equation 7–23. In applying the 'direct Fourier transform' method, one simply uses a discrete summation, in analog to Eq. 7–3. In the FFT method, w-terms need to be inserted into Eq. 7–13, defining the gridding operation; a 3–D FFT yields a three-dimensional discretely sampled image[8]; and one interpolates this result to obtain data over a spherical cap, a portion of the surface $(l,m,\sqrt{1-l^2-m^2})$. Because usually the importance of the curvature effect is minor and the data cover a small range of w, N_n, the number of slices required in the w- and n-dimensions, is small—typically eight to sixteen. The three dimensional imaging problem will be examined later in more detail in Lecture 19.

4.4. The Problem with I^D—Sidelobes

An astronomer is seldom satisfied with the approximation to I defined by I^D, or with the computed version thereof, $\widetilde{I}^D_c$. This is because of the sidelobes which contaminate I^D. As you have seen, these are due to the finite extent of the (u,v) coverage and to gaps in the coverage. Sidelobes from bright features within an image are likely to obscure any fainter features. The process described

[8] In the FFT method, one normally would want a shift of origin, in order to get the plane tangent to the celestial sphere at $(0,0,1)$ shifted to the origin of the third coordinate axis of the grid. This involves multiplying the data by $e^{2\pi i w}$, which cancels the multiplication by $e^{-2\pi i w}$ in Equation 7–23.

here is usually just the first step in obtaining a better approximation to I. Because the convolution relation $\widetilde{I}_c^D = \widetilde{B}_c^D * I$, is approximately valid, this first step provides a starting point for the deconvolution (i.e., sidelobe removal) process described in Lecture 8. However, in cases of very low signal-to-noise ratio (as might occur in an observation to determine the detectability of a putative source) one would often choose not to proceed any further. This is often the case in spectral line observing, where narrow bandwidths lead to low signal-to-noise ratios. With modern computers it is very rare to avoid deconvolution for reasons of computational capacity alone. Basic deconvolution will be described in Lecture 8, with the necessary extensions for 3–D imaging, mosaicing and multi-frequency synthesis in Lectures 19-21.

Acknowledgments. We would like to thank Alan Bridle and Rick Perley for numerous helpful discussions during the preparation of this lecture. Rick Perley kindly provided the original version of Figure 7–5.

References

Briggs, D. S. 1995, *High Fidelity Deconvolution of Moderately Resolved Sources.* Ph. D. thesis, New Mexico Institute of Mining and Technology. Available via `http://www.aoc.nrao.edu/ftp/dissertations/dbriggs/diss.html`

Bracewell, R. N. 1978, *The Fourier Transform and Its Applications*, Second Edition, McGraw–Hill, New York.

Dym, H. & McKean, H. P. 1972, *Fourier Series and Integrals*, Academic Press, New York.

Erickson, W. C., Mahoney, M. J., & Erb, K. 1982, *ApJS*, 50, 403–420.

Greisen, E. W. 1979, VLA Scientific Memorandum No. 131, NRAO.

Schwab, F. R. 1978, VLA Scientific Memorandum No. 129, NRAO.

Schwab, F. R. 1980, VLA Scientific Memorandum No. 132, NRAO.

Schwab, F. R. 1984, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 333–346.

Stratton, J. A. 1935, *Proc. Nat. Acad. Sci. U.S.A.*, 21, 51–56.

Thompson, A. R. & Bracewell, R. N. 1974, *AJ*, 79, 11–24.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

8. Deconvolution

Tim Cornwell
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Robert Braun
NFRA, Dwingeloo, The Netherlands

Daniel S. Briggs
NCSA, Champaign-Urbana, IL 61801, U.S.A.

Abstract. This lecture describes how the visibility samples collected by an interferometric array can be used to produce a high quality image of the sky. In contrast to the linear methods of Lecture 7, these methods are all non-linear, and must create estimates of the visibility function at positions in the Fourier plane where it is not measured. The two most common algorithms used, CLEAN and MEM, are discussed in detail, with several variations and less common algorithms mentioned briefly. An example comparison between CLEAN and MEM is given on a simulated VLBA observation.

1. Deconvolution

As described in Lectures 1 and 2, an interferometric array provides samples of the complex visibility function of the source at various points in the (u, v) plane. Under various approximations, which are valid for a sufficiently small source in an otherwise blank region of sky (see Lecture 1, Sec. 4.2 and Lecture 2, Sec. 6), the visibility function $V(u, v)$ is related to the source intensity distribution $I(l, m)$ (multiplied by the primary beam of the array elements) by a two-dimensional Fourier transform:

$$V(u,v) = \int\int_S I(l,m) e^{-2\pi i(ul+vm)}\, dl\, dm\,, \tag{8–1}$$

where S denotes taking the integral over the whole sky, as in Equation 2–5.

Since only a finite number of noisy samples of the visibility function are measured in practice, $I(l, m)$ itself cannot be recovered directly. Either a model with a finite number of parameters, or some stable non-parametric approach, must be used to estimate $I(l, m)$. A convenient general purpose model $\widehat{I}$ of the source intensity that is capable of representing all the visibility data consists of a two-dimensional grid of δ-functions with strengths $\widehat{I}(p\Delta l, q\Delta m)$, where Δl and Δm are the separations of the grid elements in the two orthogonal sky coordinates. The visibility $\widehat{V}$ predicted by this model is given by

$$\widehat{V}(u,v) = \sum_{p=1}^{N_l}\sum_{q=1}^{N_m} \widehat{I}(p\Delta l, q\Delta m) e^{-2\pi i(pu\Delta l + qu\Delta m)}\,. \tag{8–2}$$

For simplicity we will henceforth denote the discrete form $\widehat{I}(p\Delta l, q\Delta m)$ by the notation $\widehat{I}_{p,q}$. Assuming reasonably uniform sampling of a region of the (u, v) plane, one can expect to estimate source features with widths ranging from $\mathcal{O}(1/\max(u, v))$ up to $\mathcal{O}(1/\min(u, v))$. The grid spacings, Δl and Δm, and the number of pixels on each axis, N_l and N_m, must allow representation of all these scales. In terms of the range of (u, v) points sampled, the requirements are $\Delta l \leq \frac{1}{2u_{\max}}$, $\Delta m \leq \frac{1}{2v_{\max}}$, $N_l \Delta l \geq \frac{1}{u_{\min}}$, and $N_m \Delta m \geq \frac{1}{v_{\min}}$. This model has $N_l N_m$ free parameters, namely the cell flux densities $\widehat{I}_{p,q}$. The measurements constrain the model such that at the sampled (u, v) points

$$V(u_k, v_k) = V\widehat{(u_k, v_k)} + \epsilon(u_k, v_k)\,, \tag{8–3}$$

where $\epsilon(u_k, v_k)$ is a complex, normally distributed random error due to receiver noise, and k indexes the samples. At points in the (u, v) plane where no sample was taken, the transform of the model is free to take on any value. One can think of Equation 8–3 as a multiplicative relation

$$V(u, v) = W(u, v)(\widehat{V}(u, v) + \epsilon(u, v))\,, \tag{8–4}$$

where $W(u, v)$ is a weighted sampling function (see Lecture 7, Eq. 7–8) which is non-zero only for sampled points of the (u, v) plane,

$$W(u, v) = \sum_k W_k \delta(u - u_k, v - v_k)\,. \tag{8–5}$$

By the convolution theorem , this translates into a convolution relation in the image plane:

$$I^D_{p,q} = \sum_{p',q'} B_{p-p',q-q'} \widehat{I}_{p',q'} + E_{p,q}\,, \tag{8–6}$$

where

$$I^D_{p,q} = \sum_k W(u_k, v_k)\ \mathrm{Re}\left(V(u_k, v_k) e^{2\pi i(p u_k \Delta l + q v_k \Delta m)}\right) \tag{8–7}$$

and

$$B_{p,q} = \sum_k W(u_k, v_k)\ \mathrm{Re}\left(e^{2\pi i(p u_k \Delta l + q v_k \Delta m)}\right)\,. \tag{8–8}$$

$E_{p,q}$ in Equation 8–6 is the noise image obtained by replacing V in Equation 8–7 by $\epsilon(u_k, v_k)$. Note that the $B_{p,q}$ given by Equation 8–8 is the point spread function (beam) that is synthesized after all weighting has been applied (and after gridding and grid correction if an FFT was used; to keep the notation concise, we will not signify this gridding and grid correction explicitly). The Hermitian nature of the visibility has been used in this rearrangement.

Equation 8–4 represents the constraint that the model $\widehat{I}_{p,q}$, when convolved with the point spread function $B_{p,q}$ (also known as the *dirty beam*) corresponding to the sampled and weighted (u, v) coverage, should yield $I^D_{p,q}$ (known as the *dirty image*).

The weighting function $W(u, v)$ can be chosen to favor certain aspects of the data. For example, setting $W(u_k, v_k)$ to the reciprocal of the variance of the error in $V(u_k, v_k)$ will optimize the signal-to-noise ratio in the final image, whereas setting it to the reciprocal of some approximation to the local density of samples will minimize the sidelobe level. Robust weighting is a hybrid approach which attempts to achieve a good balance between these criteria, (see Lecture 7).

We shall now examine the possible solutions of the convolution equation.

1.1. The “principal solution” and “invisible distributions”

Let us now consider whether the convolution equation has a unique solution. Clearly if some of the spatial frequencies allowed in the model are not present in the data, then changing the amplitudes of the corresponding sinusoids in I will have no effect on the fit to the data. In effect, the dirty beam filters out these spatial frequencies. Let Z be an intensity distribution containing only these unmeasured spatial frequencies. Then $B * Z = 0$. Hence, if I is a solution of the convolution equation, so too is $I + \alpha Z$ where α is any number. Thus, as usual, the existence of homogeneous solutions implies the general non-uniqueness of any solution in the absence of boundary conditions. An important point to note is that Equation 8–6 cannot be solved by linear methods, such as $I' = A \times D$ where A is some matrix, since the homogeneous solutions Z will also be absent from I'. Thus, conventional deconvolution procedures such as inverse filtering, Wiener filtering, etc. (e.g., Andrews & Hunt 1977) will not work: a nonlinear procedure is required.

Interferometrists call the homogeneous solutions “invisible distributions” (Bracewell & Roberts 1954) or “ghosts”. The solution having zero amplitude in all the unsampled spatial frequencies is usually called the “principal” solution. Invisible distributions arise from two causes: firstly, the (u, v) coverage extends only up to finite spatial frequencies, so that the invisible distributions correspond to finer detail than can be resolved; secondly, holes may exist in the (u, v) coverage.

The problem of image construction thus can be reduced to that of choosing plausible invisible distributions to be merged with the principal solution. The shortcomings of the principal solution must be considered before tackling this problem.

1.2. Problems with the principal solution

If the data are obtained on a regular grid then the principal solution can be computed very easily: one must simply choose the weighting function in Equation 8–7 so that the bias in weight due to the vagaries of sampling are corrected. For each grid point the visibility samples are summed with appropriate weights, and the total weight normalized to unity. In such circumstances, known as uniform weighting, the principal solution is thus equal to the dirty image and is given by the convolution of the true brightness distribution with the dirty beam. For most synthesis arrays currently in use, the dirty beam has sidelobes in the range 1% to 10%. Sidelobes represent an unavoidable confusion over the true distribution of any emission in the dirty image, which can be resolved only either by making further observations or by introducing *a priori* information such as the limits in extent of the source. For example, consider uniformly weighted observations of a point source: the dirty image is just the dirty beam centered on the point source position. Without *a priori* information we cannot tell whether the source is a point or is shaped like the dirty beam. Of course we know that Stokes parameter I must be positive and that usually radio sources do not resemble dirty beams (in particular they do not have sidelobe patterns extending to infinity) and so we could use this information as an extra clue. One further unsatisfactory aspect of the principal solution, besides its implausibility, is that it changes (sometimes drastically) as more visibility data are added. A better estimator would possess greater stability.

A priori information is thus the key; in the rest of this lecture we consider two algorithms which use different constraints on the invisible distributions to derive solutions to the convolution equation. These algorithms, CLEAN and the Maximum Entropy Method (MEM), are still the predominant ones used for deconvolution of radio synthesis images.

2. The CLEAN Algorithm

The CLEAN algorithm, which was devised by J. Högbom (1974), provides one solution to the convolution equation by representing the radio source by a number of point sources in an otherwise empty field of view. A simple iterative approach is employed to find the positions and strengths of these point sources. The final deconvolved image, usually known as the CLEAN image, is the sum of these point components convolved with a CLEAN beam, usually Gaussian, to de-emphasize the higher spatial frequencies which are usually spuriously extrapolated.

We now describe some of the currently available CLEAN algorithms, including two variants of the Högbom algorithm which are better suited to large images.

2.1. The Högbom algorithm

This algorithm proceeds as follows:

1. Find the strength and position of the peak (i.e., of the point brightest in absolute intensity) in the dirty image, $I^D_{p,q}$. If desired, one may search for peaks only in specified areas of the image, called *CLEAN windows*.

2. Subtract from the dirty image, at the position of the peak, the dirty beam B multiplied by the peak strength and a damping factor γ (≤ 1, usually termed the *loop gain*).

3. Record the position and magnitude of the point source subtracted in a model.

4. Go to (1) unless any remaining peak is below some user-specified level. The remainder of the dirty image is now termed the residuals.

5. Convolve the accumulated point source model $\widehat{I}_{p,q}$ with an idealized CLEAN beam (usually an elliptical Gaussian fitted to the central lobe of the dirty beam).

6. Add the residuals of the dirty image to the CLEAN image formed in (5).

The last stage is not always performed but can often provide useful diagnostic information, for example about the noise on the map, residual sidelobes, "bowls" near the center of the image (Sec. 3.3 below), etc.

2.2. The Clark algorithm

Clark (1980) has developed an FFT-based CLEAN algorithm. A large part of the work in CLEAN is involved in shifting and scaling the dirty beam; since this is essentially a convolution it may, in some circumstances, be more efficiently performed via two-dimensional FFTs. Clark's algorithm does this, finding approximate positions and strengths of the components via CLEAN using only a small patch of the dirty beam.

In detail, the Clark algorithm has two cycles, the major and minor cycles. The *minor cycle* proceeds as follows:

1. A beam patch (a segment of the discrete representation of the beam) is selected to include the highest exterior sidelobe.

2. Points are selected from the dirty image if they have an intensity, as a fraction of the image peak, greater than the highest exterior sidelobe of the beam.

3. A list-based Högbom CLEAN is performed using the beam patch and the selected points of the dirty image. The stopping criterion for the CLEAN is roughly such that any remaining points would not be selected in step (2).

The algorithm then proceeds to a *major cycle* in which the point source model found in the minor cycle is transformed via an FFT, multiplied by the weighted sampling function that is the inverse transform of the beam, transformed back and subtracted from the dirty image. Any errors introduced in a minor cycle because of the beam patch approximation are, to some extent, corrected in subsequent minor cycles.

2.3. The Cotton–Schwab algorithm

Cotton & Schwab (as described in Schwab 1984) have developed a variant of the Clark algorithm in which the major cycle subtraction of CLEAN components is performed on the *ungridded* visibility data. Aliasing noise and gridding errors can thus be removed provided that the inverse Fourier transform of the CLEAN components to each (u, v) sample has sufficient accuracy. Two routes are used for the inverse transform: for small numbers of CLEAN components, a 'direct Fourier transform' is performed and so the accuracy is limited by the precision of the arithmetic. In the other extreme of a large number of CLEAN components, an FFT is more efficient but inevitably some errors are introduced in interpolating from the grid to each (u, v) sample. Currently, high order Lagrangian interpolation is used.

The other considerable advantage of the Cotton–Schwab algorithm, besides gridding correction, is its ability to image and CLEAN many separate but proximate fields simultaneously. In the minor cycle each field is CLEANed independently, but in the major cycles, CLEAN components from all fields are removed. In calculating the residual image for each field, the full phase equation, including the w-term, can be used. Thus, the algorithm can correct what is commonly called the "non-coplanar baselines" distortion of images (see Lectures 2, 7 and 19).

The Cotton–Schwab algorithm is often faster than the Clark CLEAN, the major exception occurring for data sets with a large number of visibility samples, where gridding over and over again becomes prohibitively expensive. The Cotton–Schwab algorithm also allows CLEANing with smaller guard bands around the region of interest, hence with smaller image sizes.

This algorithm is implemented in NRAO's Astronomical Image Processing System (AIPS) as the classic task `MX` and the modern version, `IMAGR`.

2.4. Other related algorithms

Several algorithms have been invented with the aim of correcting some deficiencies of CLEAN.

Steer, Dewdney & Ito (1984) developed a variant of the Clark algorithm in which the minor cycle is replaced by a step of simply taking all points above a sidelobe-dependent threshold, scaling them and then subtracting normally in the major cycle. The saving in time can be considerable compared to CLEAN, but the radio astronomy community has relatively little experience with this variant of the algorithm. For some sources it can suppress the well known striping of extended emission, but for high precision deconvolution of moderately compact objects, it does not appear superior to the basic CLEAN.

Segalovitz & Frieden (1978) proposed an *ad hoc* modification of the *dirty* beam to enhance the smoothness of the resulting CLEAN image. Cornwell (1983) justified a similar prescription as forcing the minimization of the image power (i.e., the sum of the squares of the pixel values) and thus pushing down the extrapolated visibility function. Both approaches seem again to ameliorate partially the striping instability to which CLEAN is susceptible but possibly at cost in the overall stability of the algorithm.

Keel (1988) extended the domain of CLEAN to conventional optical imaging with the 'σ-CLEAN', where instead of searching for the maximum in the dirty image residuals, one searches for the peak in signal-to-noise at each pixel. This change is necessary due to the different character of the noise involved. In spite of working quite adequately, this has not proved popular in optical work and other restoration algorithms are generally used instead.

2.5. Practical Details and Problems of CLEAN Usage

Theoretical understanding of CLEAN is relatively poor even though the original algorithm is about 25 years old. Schwarz (1978, 1979) has analyzed the Högbom CLEAN algorithm in some detail. He notes that in the noise-free case the least-squares minimization of the difference between observed and model visibility, which CLEAN performs, produces a unique answer if the number of cells in the model is not greater than the number of independent visibility measurements contributing to the dirty image and beam (*cf.* Eqs. 8–7 and 8–8), counting real and imaginary parts separately. This rule is unaffected by the distribution of (u, v) sample points so that, in principle, super-resolution is possible if enough data points are available. In practice, however, the introduction of noise and the use of the FFT algorithm to calculate the dirty image and beam corrupts our knowledge of the derivatives of the visibility function upon which super-resolution is based. (As shown in Chapter 5 of Briggs (1995), CLEAN is in fact particularly *bad* at the visibility extrapolation involved in super-resolution and is not recommended for the purpose.) Even if the FFT is not used, the presence of noise means that independence of the data must be redefined. Schwarz has in fact produced a noise analysis of the least-squares approach but it involves the inversion of a matrix of side $N_l N_m$ and so is totally impractical for typical image sizes; furthermore, we are really interested in CLEAN, not the more limited least-squares method, since CLEAN will still produce a unique answer in circumstances where the least-squares method is guaranteed to fail. To date no one has succeeded in producing a noise analysis of CLEAN itself. The existence of instabilities in CLEAN, which will be discussed later, makes such an analysis highly desirable.

Schwarz also proves three conditions for the convergence of CLEAN:

1. The beam must be symmetric.

2. The beam must be positive definite or positive semi-definite. Thus the eigenvalues must be non-negative.

3. The dirty image must be in the *range* of the dirty beam. Roughly speaking, there must be no spatial frequencies present in the dirty image which are not also present in the dirty beam.

All three of these conditions are obeyed in principle for the dirty image and beam calculated by Equations 8–7 and 8–8 if the weighting function is nowhere negative. In practice, however, numerical errors, and the gridding and grid-correction process may cause violation of these conditions. The CLEAN algorithm will therefore diverge eventually. CLEANing close to the edge of a dirty image computed by an FFT is particularly risky. Even the most simple case of a three pixel image has been demonstrated by Tan (1986) to be potentially chaotic. Still, in real cases the algorithm seems to work well.

Marsh & Richardson (1987) showed that for the case of an isolated point source, the CLEAN algorithm approximately minimizes the sum of the pixel values in the component model, subject to the constraint that these are positive. That is, it returns the minimum flux solution consistent with the data. By comparison with empirical results of deconvolvers which explicitly minimize this criterion, clearly this is only an approximation even in simple cases—it is not obvious how far this insight should be trusted.

Thus most of our understanding of CLEAN still comes from a combination of guessing how to apply intuition and Schwarz's analysis to real cases, and much practical experience on real and test data. In the rest of this section we will attempt to summarize the current lore concerning how the algorithm should be used, and how it can fail.

2.6. The use of boxes

The region of the image which is searched for the peak can be limited to those areas (known as the CLEAN *windows* or *boxes*) within which emission is known or guessed to be present. These boxes effectively restrict the number of degrees of freedom available in the fitting of the data. Schwarz's work (and common sense) tells us that the number of such degrees of freedom should be minimized but that the CLEAN window should include all real emission in the image. For a simple source in an otherwise uncluttered field of view, one CLEAN window will do, but multiple boxes may be needed when CLEANing more complicated sources, or for a field containing many sources. In the latter case, the presence of weak sources may be revealed only after the sidelobes of the stronger sources have been removed; more boxes may therefore be required as the CLEAN progresses. Note that such *a posteriori* definition of CLEAN boxes considerably complicates any possible noise analysis.

The practical implications of Schwarz's observation that the number of degrees of freedom should not exceed the number of independent constraints are difficult to gauge. In the presence of noise (u, v) points should be judged independent if the differences in visibility due to the size of structure expected are much greater than the noise level. Counting visibility points in such a way, the aggregate area of the CLEAN boxes in pixels should be less than twice the number of *independent* visibility points. If the FFT is used (see Lecture 7) then the number of independent visibility samples cannot be greater than $\mathcal{O}(N_l N_m)$, and so the use of CLEAN boxes is certainly advisable.

Given the uncertainty in determining the number of independent data points, and hence the number of constraints, caution dictates that boxes should always be placed tightly around the region to be CLEANed.

2.7. Number of iterations, loop gain and the beam patch size

The number of CLEAN subtractions N_{CL} and the loop gain γ determine how deep the CLEAN goes. In particular, for a point source the residual left on the dirty image is $(1-\gamma)^{N_{CL}}$. Hence, to minimize the number of CLEAN subtractions (and so to minimize the CPU time) γ should be unity; one then finds, however, that extended structure is not well represented in the corresponding CLEAN image. In typical VLA applications a reasonable compromise lies in the range $0.1 \leq \gamma \leq 0.25$. (Incidentally, this dependence of the CLEAN image upon the loop gain is a nice demonstration of the multiplicity of solutions to the convolution equation.) Lower loop gains may be required in cases where the (u, v) coverage is poor, but experience suggests that the improvements in deconvolution for $\gamma \ll 0.01$ are generally minimal. If one is in any doubt then it is wise to experiment (e.g., by decreasing γ and increasing N_{CL}). One exception to the use of low loop gain is in the removal of confusing sources; it is preferable to remove them with high loop gain, as their structure is usually not of interest.

The choice of the number of iterations depends upon the amount of real emission in the dirty image. One should aim at transferring all brightness greater than the noise level to CLEAN components (some implementations of CLEAN allow one to specify a lower intensity limit to the components instead of N_{CL}). CLEANing deep into the noise is usually a waste of time unless you specifically wish to analyze the extended, low surface-brightness emission. For high dynamic range imaging, the highest deconvolution fidelity will occur when CLEANing very deeply, but the very act of CLEANing noise will alter the statistics and risks making the image appear better than it really is.

Examination of the list of CLEAN components, and, in particular, of the behavior of the accumulated intensity in the model, is useful in detecting divergence; sometimes the accumulated intensity diverges. As discussed above, divergence of the Högbom CLEAN is always due to a computational problem. Possible culprits are the gridding process, aliasing, and finite precision arithmetic. In the case of the Clark or the Cotton–Schwab algorithms, the truncated dirty beam patch that is used in the minor cycles of these algorithms must violate Schwarz's conditions. Therefore both may be subject to instability or divergence if the minor cycle is prolonged unduly. The default size of the beam patch in deconvolution programs was often set at a time when computer memory was at more of a premium than it is today. For large sources, it is often a wise idea to override this choice and use a larger beam patch than the default. In addition to obviously helping to delay divergence in the minor cycle it can also improve the overall level of the deconvolution by ensuring that there are smaller errors to be corrected after each major cycle than otherwise.

2.8. The problem of short spacings

Implicit in deconvolution is the interpolation of values for unsampled (u, v) spacings. In most cases CLEAN does this interpolation reasonably well. However, in the case of short spacings the poor interpolation is sometimes rather more noticeable since very extended objects have much more power at the short spacings. The error is nearly always an underestimation and is manifested as a "bowl" of negative surface-brightness in which the source rests. In such a case, introducing an estimate of the zero-spacing flux density into the visibility data before forming the dirty image will sometimes help considerably. The appropriate value of this flux density would be that measured by a single element of the array. In practice, however, single array elements rarely have sufficient sensitivity or stability to provide this estimate accurately. Values estimated from surveys made with larger, more sensitive, and more directive elements are therefore frequently substituted. Choosing the weight for the zero-spacing flux density is difficult; the best estimate seems to be simply the number of unfilled cells around the origin of the gridded (u, v) plane. However, the results obtained are fairly insensitive to the value used *provided that the CLEAN deconvolution goes deep enough.*

The CLEAN windows or boxes may also be viewed as providing crude estimates of the shape of the visibility function near the zero spacing $u = v = 0$. For this reason, careful choice of CLEAN windows may also minimize problems associated with the short spacings.

After CLEANing, the emission should be, but is not guaranteed to be, distributed sensibly over the CLEAN image. Failure of the interpolation is indicated by the presence of a "pedestal" of surface brightness within the CLEAN box upon which the source rests. Such a pedestal all over the CLEAN image can be caused by insufficient CLEANing of the dirty image; one can experiment by simply increasing N_{CL}. Ultimately, it may actually be necessary to measure the appropriate data!

2.9. The CLEAN beam

The CLEAN beam (more generally called the *restoring beam*) is used to suppress the higher spatial frequencies which are poorly estimated by the CLEAN algorithm. There are two competing opinions on this in the radio astronomy community: some object that it is purely *ad hoc* and is undesirable—in the sense that the equivalent predicted visibilities do not then agree with those observed. Others defend it as a way of recognizing the inherent limit to resolution.

In practice, it does appear to be necessary in order to produce astrophysically reasonable images.

The magnitude of just how poorly CLEAN extrapolates past the sampling envelope has only recently been appreciated—the errors in the restored image comes almost always from the extrapolated region in the (u, v) plane and rarely from interior holes in the sampling. Thus there is a straightforward tradeoff in resolution against image fidelity controlled by the size of the restoring beam. The most common method of choosing the restoring beam is to fit an elliptical Gaussian to the central region of the dirty beam, but this default is not mandated. A smaller restoring beam allows more of the erroneously extrapolated model into the final solution and yields poorer fidelity in the name of higher resolution. Conversely, a larger than default CLEAN beam can produce a highly accurate deconvolution. This tradeoff is explored pragmatically in Chapter 5 of Briggs (1995).

Various attempts have been made to improve the selection of the CLEAN beam. The dirty beam, truncated outside the first zero-crossing, is appropriate in some applications since it lacks the extended wings of a Gaussian, but we emphasize that, after convolution with such a beam, the CLEAN image does not agree satisfactorily with the original visibilities. An ideal CLEAN beam might be defined as a function obeying three constraints:

1. Its transform should be unity inside the sampled region of the (u, v) plane.

2. Its transform should tend to zero outside the sampled region as rapidly as possible.

3. Any negative sidelobes should produce effects comparable with the noise level in the CLEAN image.

Constraint (1) is usually the first to be relaxed, and then only positivity of the transform is necessary. It may be that in typical applications CLEAN performs so poorly that these constraints do not allow an astrophysically plausible CLEAN image, however such a topic is probably worth further consideration.

One very important consequence of a poor choice for the CLEAN beam is that the units of the convolved CLEAN components may not agree with the units of the residuals. The units of a dirty image are not very well defined but can be called "Jy per dirty beam area". The only real meaning of these units is that an isolated point source of flux density S Jy will show up in the dirty image as a dirty beam shape with amplitude S Jy per dirty beam area. An extended source of total flux density S Jy will be seen in the dirty image convolved with the dirty beam, but the integral will not, in general, be S Jy. However, convolved CLEAN components do have sensible units of Jy per CLEAN beam, which can be converted to Jy per unit area since the equivalent area of the CLEAN beam is known. Careful control of the dirty beam shape with weighting parameters as described in Lecture 7 can often produce a more Gaussian-like dirty beam than the typical defaults, resulting in a better match in the flux scale between convolved components and residuals. A few imaging programs will also rescale the residuals with the ratio between the CLEAN beam and the fitted beam before adding these to the convolved components, but neither approach is a perfect correction. This issue is most important when significant flux remains in the residuals, so when using very non-Gaussian dirty beams and/or a non-fitted restoring beam, it is best to run CLEAN quite deeply and transfer as much flux density to the components as possible. In this limit, the integral of the CLEAN image will often provide an accurate estimate of the flux density of an extended object—surprisingly often better than that of MEM—usually failing when the u-v coverage is incomplete on the spacings required. If convergence is not attained then both flux density and noise estimates taken from the CLEAN image can be in error.

2.10. Use of *a priori* models

A priori models of sources can be used to good effect in CLEAN. Perhaps the best example is in the CLEANing of images of planets; in this case the visibility function of a circular disk can be subtracted from the observed visibilities before making the dirty image. CLEAN then needs only to find the small perturbations from the disk model, and so both the image quality and speed of convergence should be improved.

2.11. Non-uniqueness

Perhaps the biggest drawback to the use of CLEAN is the way in which the answers depend upon the various control parameters: the CLEAN boxes, the loop gain and the number of CLEAN subtractions. By changing these one can, even for a relatively well-sampled (u, v) plane, produce somewhat different final CLEAN images. In the absence of an error analysis of CLEAN itself one can do nothing at all about this problem. Awareness of the possible effects discussed in this section should however keep you from becoming over-confident in the final CLEAN image, as will experience of applying CLEAN to a wide range of different images.

In any one application, Monte Carlo tests of CLEAN can sometimes be illuminating, and, indeed, provide the only means of estimating the effects of various data errors and CLEANing strategies upon the final image.

2.12. Instabilities

One particular instability of CLEAN is well known: in CLEAN images of extended sources one sometimes finds modulations at spatial frequencies corresponding to unsampled parts of the (u, v) plane (see, e.g., Cornwell 1983 for an example). Convolution with a larger than usual CLEAN beam will sometimes mask this problem, especially when the unsampled region is in the outer parts of the (u, v) plane. Reducing the loop gain γ to very low values generally has little effect, but there is reason to believe that the instability is triggered by noise and hence that *temporarily* setting the loop gain equal to the noise-to-signal ratio when the instability begins may help (U. J. Schwarz, private communication).

Cornwell (1983) has developed a simple modification to the CLEAN algorithm that is sometimes successful in countering the instability. A small-amplitude delta function is added to the peak of the beam before CLEANing. The effect of the spike is to perform negative feedback of the CLEAN structure into the dirty image, and thus to act against any features not required by the data. Spike heights of a few percent, and lower loop gains than usual are usually required. In view of the limited success of this modification, a better solution is to use another deconvolution algorithm, such as MEM.

The occurrence of the stripes is a natural consequence of the incorrect information about radio sources embodied in the CLEAN algorithm. Astronomers very rarely find convincing evidence for the existence of such stripes in radio sources and so they are skeptical about such stripes when found in CLEAN images. Unfortunately the only *a priori* information built into CLEAN, via the use of CLEAN boxes, is that astronomers prefer to see mainly blank images; there is no bias against stripes. Such considerations, and some others, have led to the development of deconvolution algorithms which either incorporate extra constraints on astrophysically plausible brightness distributions or are claimed to produce, in some way, optimal solutions to the deconvolution equation. In the next section we briefly consider one such algorithm.

3. The Maximum Entropy Method (MEM)

The deconvolution problem is one of selecting one answer from the many possible. The CLEAN approach is to use a *procedure* which selects a plausible image from the set of feasible images. Some of the problems with CLEAN arise because it is procedural so that there is no simple equation describing the CLEAN image. Thus, for example, a noise analysis of CLEAN is very difficult. By contrast, the Maximum Entropy Method (MEM) is not procedural: the image selected is that which fits the data, to within the noise level, and also has maximum entropy. The use of the term *entropy* has lead to great confusion over the justification for MEM. Even today there is no consensus on this subject evident in the literature (e.g., Frieden 1972; Wernecke & D'Addario 1976; Gull & Daniell 1978; Jaynes 1982; Narayan & Nityananda 1984, 1986; Cornwell & Evans 1985). We will use the "lowest common denominator" justification and define entropy as something, which when maximized, produces a positive image with a compressed range in pixel values. Image entropy is therefore not to be confused with a "physical entropy", although the logarithmic definition given in equation 8-9 parallels that of of the Boltzman H-function in statistical mechanics (see Cornwell 1984, Landau & Lifshitz 1980). The compression in pixel values forces the MEM image to be "smooth", and the positivity forces super-resolution on bright, isolated objects. There are many possible forms of this extended type of entropy, see e.g., Narayan & Nityananda 1984, but one of the best for general purpose use is:

$$\mathcal{H} = -\sum_k I_k \, \ln \frac{I_k}{M_k e}\,, \tag{8–9}$$

where M_k is a "default" image incorporated to allow *a priori* knowledge to be used. For example, a low resolution image of the object can be used to good effect as the default.

A requirement that each visibility point be fitted exactly is nearly always incompatible with the positivity of the MEM image. Consequently, data are usually incorporated in a constraint that the fit, χ^2, of the predicted visibility to that observed, be close to the expected value:

$$\chi^2 = \sum_k \frac{|V(u_k, v_k) - \widehat{V(u_k, v_k)}|^2}{\sigma^2_{V(u_k,v_k)}} . \tag{8–10}$$

Simply maximizing $\mathcal{H}$ subject to the constraint that χ^2 be equal to its expected value leads to an image which fits the long spacings much too well (better than 1σ) and the zero and short spacings very poorly. The cause of this effect is somewhat obscure but is related to the fact that the entropy $\mathcal{H}$ is insensitive to spatial information. It can be avoided by constraining the predicted zero-spacing flux density to equal that provided by the user (Cornwell & Evans 1985).

Algorithms for solving this maximization problem have been given by Wernecke & D'Addario (1976), by Cornwell & Evans (1985), and by Skilling & Bryan (1984). The Cornwell–Evans algorithm is coded in NRAO's Astronomical Image Processing System (AIPS) as `VM` or `VTESS`. It is generally faster than CLEAN for larger images; the break-even point being for images of about 1 million pixels.

4. Practical Details of the Use of MEM

The following description relates to the AIPS MEM algorithm, `VM`.

4.1. The default image (prior distribution)

Examination of Equation 8–9 reveals that if no data constraints exist, the MEM image is the default image, so the MEM image is always biased towards the default. A reasonable "default default" image is flat, with total flux density equal to that specified. A low-resolution image, if available, can be used as the default to very good effect; this is a nice way of combining single-dish data with interferometric data. A spike in the default can sometimes be used to indicate the presence of an unresolved source, which could otherwise cause problems (see Sec. 4.5 below).

4.2. Total flux density

As described above, if the total flux density in the MEM image is not specified then the value found may be seriously biased if the signal-to-noise ratio is low. There is no real way around this at the moment, except by guessing a value and then adjusting it to get an image that looks "reasonable"—for example, possessing a flat baseline. For bright objects, only an order-of-magnitude estimate is required to set the flux density scale. Of course, then the estimated flux density is not fitted but is used only to set a reasonable default image.

4.3. Varying resolution

In the folklore, MEM is criticized for resolution that depends on the signal-to-noise ratio. In fact, there are sound theoretical reasons to believe that this effect is common to all nonlinear algorithms that know about noise (Andrews & Hunt 1977). If you want to "fix" the resolution in MEM, you basically have two choices:

1. Convolve the final MEM image with a Gaussian beam of appropriate width to smear out the fine scale structure and add the residuals back in.

2. Before deconvolution, convolve the dirty image with a Gaussian beam.

The advantages of (2) over (1) are that the algorithm usually converges faster, and that given the nonlinear nature of the deconvolution, the answer can be (and usually is) better. For example, sidelobes around a point source embedded in extended emission are not well removed by MEM, whereas scheme (2) often alleviates this effect. The advantages of (1) over (2) are that both image bias (see below) and errors in gradient representation are substantially alleviated by adding in the residuals.

There are occasions when the super-resolution exhibited by MEM images is reliable, although predicting this in advance is not feasible. With careful modeling of the source, however, it is possible to plausibly defend the physical reality of super-resolved features, as in Chapter 8 of Briggs (1995). MEM is in fact the algorithm of choice for super-resolution studies.

4.4. Bias

Another commonly heard complaint about MEM is that the answer is biased, i.e., that the ensemble average of the estimated noise is not zero. This is certainly true, and is the price paid by any method which does not try to fit exactly to the data as CLEAN does. Bias in an estimator is quite common and acceptable since it usually leads to smaller variance. Cornwell (1980) has estimated the magnitude of the bias, and has shown that it is much less than the noise for pixels having signal-to-noise ratio much greater than one. In fact, if the (u, v) coverage is very good then for bright pixels the effect of noise on an MEM image is very similar to that on a dirty image. The effect of bias can be substantially reduced by using a reasonable default such as a previous MEM image smoothed with a Gaussian; then only the highest spatial frequencies are biased. The effect of bias can also be eliminated by adding back the residuals after ensuring a similar flux scale via convolution of the MEM image with a Gaussian as outlined above.

4.5. Point sources in extended emission

Nearly all the power of MEM to remove sidelobes comes from the positivity constraint. Hence, if the source sits on a background level of emission, then the sidelobes will not be removed fully. The only consistently effective solutions are either (a) to remove the point sources using CLEAN or (b) to smooth the dirty image prior to deconvolution. MEM has difficult even with isolated point sources without a background, but only a small degree of resolution – say an intrinsic feature width of 1/5 of a beamwidth – is necessary for the algorithm to perform well.

5. Comparison of CLEAN and MEM

CLEAN has dominated deconvolution in radio astronomy since its invention nearly 25 years ago, but has not been widely applied in other disciplines. One of the major reasons for this is the decomposition into point sources, which is often not permissible in other types of images. In contrast, MEM has spread to many different fields, probably because most of the justifications are independent of the type of data to which it is applied.

The philosophy behind MEM is intriguing and may convince some of you about the objectivity of MEM (see Jaynes 1982 for an exposition of MEM from its inventor). For those of you who do not become acolytes, the practical differences between CLEAN and MEM are probably more interesting.

CLEAN is nearly always faster than MEM for sufficiently small and simple images, because its approach of optimizing a relatively small number of pixels is simply more efficient. For typical VLA images, the break-even point is at around a million pixels of brightness. For very large and complex images, such as those of supernova remnants, which may contain up to 100 million pixels, CLEAN is impossibly slow and an MEM-type algorithm is absolutely necessary.

CLEAN images are nearly always rougher than MEM images. This may be traced to the basic iterative scheme: since what happens to one pixel is not coupled to what happens to its neighbors, there is no mechanism to introduce smoothness. MEM couples pixels together by minimizing the spread in pixels' values, so the resulting images look smooth although the entropy term does not explicitly contain spatial information.

Both MEM and CLEAN fail to work well on certain types of structure. CLEAN usually makes extended emission blotchy, and may introduce coherent errors such as stripes, while MEM copes very poorly with point sources in extended emission. Both work quite well on isolated sources with simple structure, and can produce meaningful enhancement of resolution, although MEM seems to do better in most cases.

Since MEM tries to separate signal and noise, it is necessary to know the noise level reasonably well. Also, as mentioned above, knowledge of the total flux density in the image helps considerably. Apart from this MEM has no other important control parameters, although it can be helped enormously by specifying a default image. CLEAN makes no attempt to separate out the noise, and so specification of the noise level is not required. The main control parameters are the loop gain γ, and the number of iterations N_{CL}, both of which are important in determining the final deconvolution.

The default image of MEM is a very powerful mechanism for introducing *a priori* information. We have previously described the use of a simple image as a default; however, the default image need not be only a simple fixed set of numbers, but instead can be used to introduce functional relationships between pixels. For example, to further encourage smoothness, one might make the default for a pixel equal to the geometric mean of the brightness of its neighbors (S. F. Gull, private communication). Only the simple fixed default image can be easily mimicked by CLEAN: the default image is simply used as the starting point for the collection of CLEAN components. Thus the use of a disk model for a planet is an example of the use of a default in CLEAN.

6. Example

Figures 8–1 and 8–2 give an example deconvolution of a core jet source, adapted from Cornwell (1995). The data are synthetic, with the model resembling M87 and scaled to the VLBI size regime. The source was 'observed' with the VLBA at a frequency of 1.6 GHz, the declination was 50°, and the coverage was horizon to horizon down to 15° elevation. That is, the (u, v) coverage is superb by VLBI standards, and medium to poor by those of the VLA. The bright point source core and extended jet was designed to demonstrate the strengths and weaknesses of the two algorithms, with CLEAN performing better on the core, and MEM performing better on the extended emission. A small amount of thermal noise was added – below the lowest contour level – but the calibration was assumed to be perfect.

Figure 8–1 shows the parameters and truth image of the simulation. Panel **(a)** shows the model, smoothed to the same resolution as the restored images. The (u, v) coverage is in panel **(b)** and the visibility amplitude in panel **(c)**. Notice that the total flux density is dominated by the extended emission, yet the point source core will totally dominate the deconvolution in some respects. Panels **(d)** and **(e)** show the uniformly weighted beam used in this simulation. The former has lowest contour at ±2.2% and the latter is a typical slice though the central portion of the beam.

Figure 8–2 presents the results of different deconvolution strategies. Panel **(a)** is a simple Clark CLEAN with a loop gain of 0.1, run to 20,000 components without any constraint on their position. (The contours are roughly powers of two from a low of 0.05%.) The image has greatly improved from the dirty image (not shown), but there is still some evidence of sidelobes paralleling the jet. In panel **(b)**, components have only been allowed in a tight region surrounding the model source. The same 20,000 components now produce a very good image, showing the incorporation of information from the support constraints. Panel **(c)** is the same image, but contoured starting a factor of 10 lower.

The lower three panels are all MEM images. The first, panel **(d)** was generated with a flat default and the same support constraints as in panel **(b)**. 80 iterations were used, as compared to the more typical 30 and still MEM is having great difficulty with the point source core. (The image without the support constraint was even poorer.) In panel **(e)**, a point source model was fitted to the core of the CLEAN image, then subtracted from the visibility data. The residuals were imaged with MEM (and the support constraint), and then final image reassembled—the difference is dramatic. Panel **(f)** has this same image contoured down at the level of panel **(c)**. The best MEM image is smoother at the lowest contour levels than the best CLEAN image, and has a different characteristic error pattern. The images are of comparable fidelity. Remember that MEM has the most difficulty with *point* sources. If the core had been resolved so much as half a beam width, the initial MEM image would have been comparable to the CLEAN image without needing the subtraction.

7. Other Methods, Including Hybrids

Deconvolution in radio astronomy is still dominated by two *nonlinear* algorithms, CLEAN and MEM. Other nonlinear algorithms exist and may turn out to be useful, at least in the sense that, as with CLEAN and MEM, their defects are orthogonal to those of other algorithms. This property of defect orthogonality also suggests the use of a combination of algorithms in the deconvolution of a single image, so that the virtues of each approach can be exploited. More novel approaches to deconvolution are under development, but have yet to be transfered into the mainstream of mundane reduction.

With modern computers, it is now possible to solve for directly the parameters of the point source model for some interesting objects, via brute force constrained least squares optimization. Briggs (1995) has applied the NNLS algorithm of Lawson & Hanson to solve the convolution equation for compact objects. At present this seems computationally feasible for images having nonzero flux in about 5000–6000 pixels. The quality of the deconvolved images is excellent for such sources, though it is actually the interaction of the algorithm with self-calibration which might prove the most important. A weakness of both CLEAN and MEM is that they produce highly correlated error patterns in the (u, v) plane. This correlated error pattern can prove a significant problem when the deconvolved model is used as input for a self-calibration correction, (see Lecture 10). CLEAN is better than MEM in this regard, though both algorithms can cause the deconvolution/self-calibration hybrid mapping cycle to stall. By contrast, NNLS appears to yield a much flatter and less correlated error pattern in the (u, v) plane and interacts extremely well with self-calibration. This approach might prove fairly useful for VLBI and for very high dynamic range VLA applications.

Multi-resolution algorithms are becoming more attractive. These all rely implicitly on the notion that deconvolution of simple objects is easier than complicated ones. In some cases, one actually iteratively solves related deconvolution problems at different scales, such as in the multi-resolution CLEAN of Wakker & Schwarz (1991). In others, the multi-resolution aspect is reflected in the decomposition of the problem into a wavelet domain. (See, e.g., Starck et al. 1994 or Pantin & Starck 1996). The wavelet based methods in particular seem very promising, but as yet available software is still an impediment to widescale exploration of these algorithms.

Hybrid techniques attempt to exploit the virtues, while avoiding the pitfalls, of a number of algorithms simultaneously. For example, the awkward but common circumstance of deconvolving compact structure on an extended background can be successfully approached with a shallow CLEANing of compact structure down to the level of the extended emission, followed by a MEM deconvolution of what remains. The component models of each method are then combined, restored, and added to the residuals. A further variant of this approach which is also effective for multi-pointing deconvolution problems consists of CLEANing the individual pointings at the full available resolution and forming the linear combination with appropriate weighting, while using MEM to simultaneously deconvolve the data at very low resolution. These results are then merged by extracting the inner Fourier transform plane of the MEM result and combining it (with appropriate normalization) with the outer Fourier transform plane of the CLEAN result and back-transforming. Surprisingly, while these techniques have now been used successfully for many years, there is still no streamlined datapath for the hybrid approachs. The scientist must still do

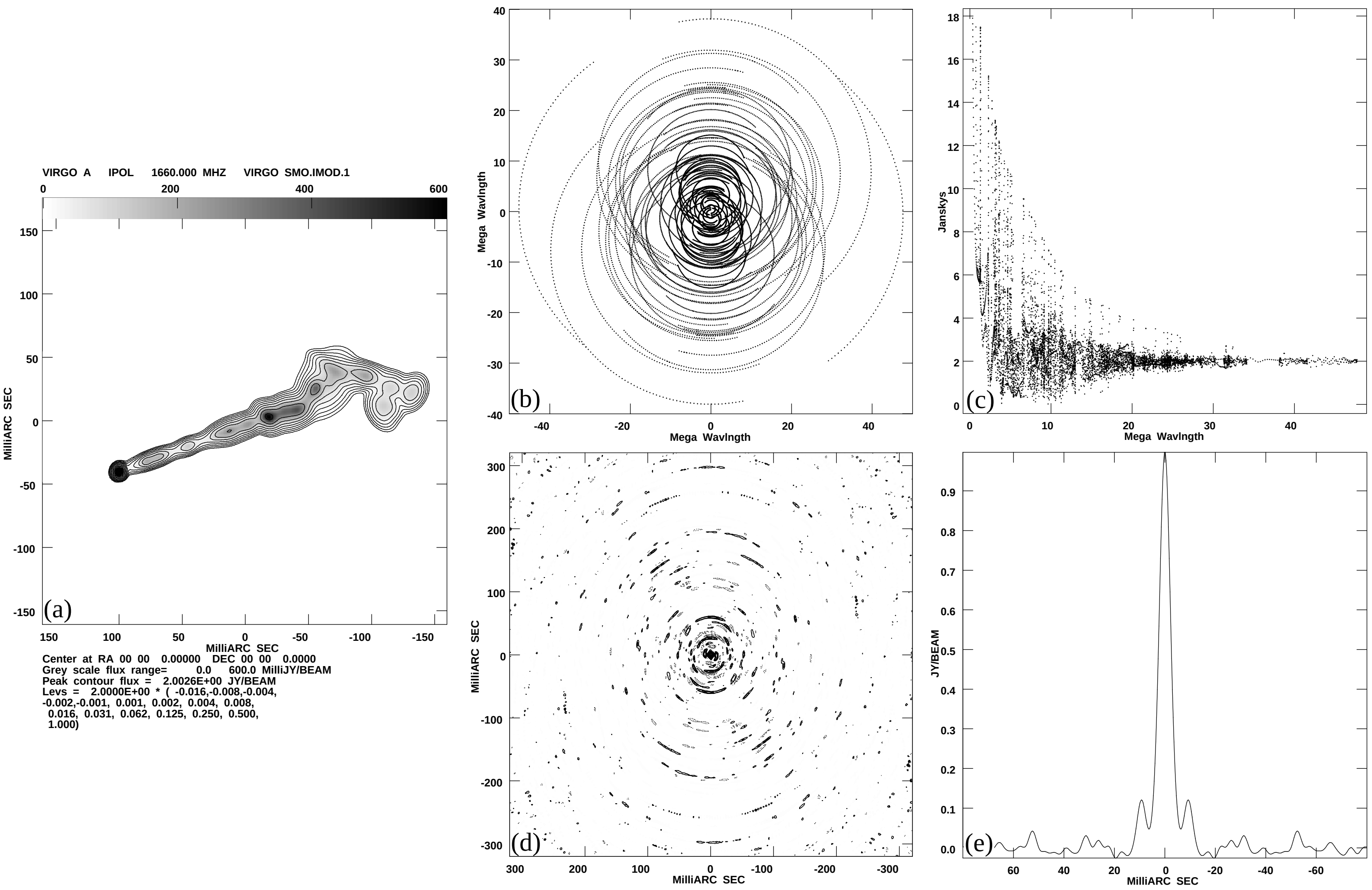


Figure 8–1. Example deconvolution: See text for details.

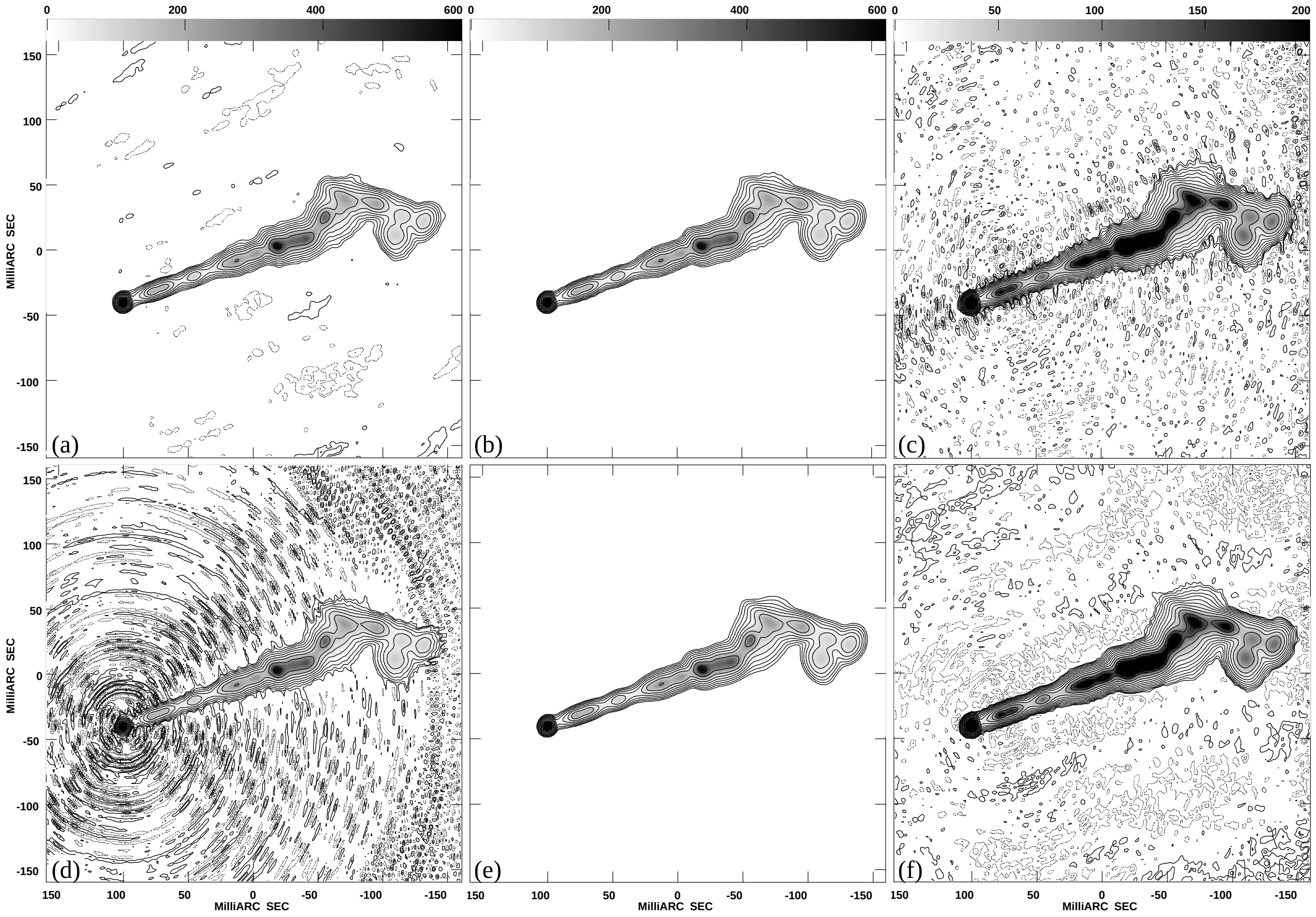


Figure 8–2. Example continued.

much of the bookkeeping involved in combining the results of the different sub-algorithms.

It is ironic that, formally, more is known about the type of images generated by MEM than by CLEAN (see e.g., Narayan & Nityananda 1986), since CLEAN is rather more widely used. Indeed many of the criticisms of MEM arise because certain of its properties, such as the bias, can be analyzed. Schwarz's analysis of CLEAN is incomplete in that it does not address the interesting underdetermined case in which there are fewer data than pixels. We hope that someday this problem might be investigated satisfactorily.

Although deconvolution algorithms are now as important in determining the quality of images produced by a radio telescope as the receivers, correlators and other equipment, they are far less well understood. A good description is that they are poorly engineered. Only further research and development of new and existing algorithms can redress this imbalance.

References

Andrews, H. C. & Hunt, B. R. 1977, *Digital Image Restoration*, Prentice–Hall (Englewood Cliffs, NJ).

Bracewell, R. N. & Roberts, J. A. 1954, *Aust. J. Phys.*, 7, 615–640.

Briggs, D. S. 1995, Ph. D. thesis, New Mexico Institute of Mining and Technology. Available via `http://www.aoc.nrao.edu/ftp/dissertations/dbriggs/diss.html`

Clark, B. G. 1980, *A&A*, 89, 377–378.

Cornwell, T. J. 1980, Ph. D. Thesis, University of Manchester.

Cornwell, T. J. 1983, *A&A*, 121, 281–285.

Cornwell, T. J. 1984, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 291–296.

Cornwell, T. J. & Evans, K. F. 1985, *A&A*, 143, 77–83.

Cornwell, T. J. 1995, in *Very Long Baseline Interferometry and the VLBA*, Zensus et al., Eds., San Francisco: Astronomical Society of the Pacific

Frieden, B. R. 1972, *J. Opt. Soc. Am.*, 62, 511–518.

Gull, S. F. & Daniell, G. 1978, *Nature*, 272, 686–690.

Högbom, J. 1974, *ApJS*, 15, 417–426.

Jaynes, E. T. 1982, *Proc. IEEE*, 70, 939–952.

Keel, W. C. 1988, *ApJ*, 329, 532–550.

Landau, L.D., & Lifshitz, E.M. 1980, *Statistical Physics*, New York: Pergamon Press, p. 119.

Marsh, K. A. & Richardson, J. M. 1987, *A&A*, 182, 174–178.

Narayan, R. & Nityananda, R. 1984, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 281–290.

Narayan, R. & Nityananda, R. 1986, *Ann. Rev. Astron. Astrophys.*, 24, 127–170.

Pantin, E. & Starck, J.-L. 1996, *A&AS*, 118, 575–585

Schwab, F. R. 1984, *AJ*, 89, 1076–1081.

Schwarz, U. J. 1978, *A&A*, 65, 345–356.

Schwarz, U. J. 1979, in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., D. Reidel (Dordrecht, Holland), pp. 261–275.

Segalovitz, A. & Frieden, B. R. 1978, *A&A*, 70, 335–343.

Skilling, J. & Bryan, R. K. 1984, *MNRAS*, 211, 111–124.

Starck, J.-L. et al., 1994 *A&A*, 283, 349–360

Steer, D. G., Dewdney, P. E., & Ito, M. R. 1984, *A&A*, 137, 159–165.

Tan, S. 1986, *MNRAS*, 220, 971–1001

Wernecke, S. J. & D'Addario, L. R. 1976, *IEEE Trans. Computers*, C-26, 351–364.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

9. Sensitivity

J. M. Wrobel and R. C. Walker
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. An introduction to sensitivity in synthesis imaging, with examples.

1. What is Sensitivity?

Sensitivity is a measure of the weakest source of radio emission that can be detected. An understanding of sensitivity is key to all stages of a research project, from preparing a technically sound observing proposal to conducting a sensible error analysis in a publication in the astronomical literature. This lecture will take a bottom-up approach to introducing you to sensitivity in synthesis imaging. The basic building block of any radio synthesis array is a radio antenna; Section 2 will discuss performance measures for an antenna. Coupling two antennas appropriately builds a radio interferometer; Section 3 uses the results of Section 2 to describe the sensitivity of an interferometer, expressed in units of Janskys [1]. Finally, a radio synthesis image contains information from many two-element interferometers, so Section 4 will build on the results of Section 3 to describe the sensitivity of an image, expressed in units of Janskys per synthesized beam area. Practical examples will be tossed in along the way.

2. What Are Antenna Performance Measures?

Radio astronomers find it convenient to refer to the power of various signals from a radio telescope in terms of the equivalent temperature, T, of a matched termination on the input of the receiver (Crane & Napier 1989; Walker 1995). Using the Rayleigh-Jeans approximation to the Planck radiation law for a black body, such a power is given by:

$$P = k_{\rm B}\, T\, \Delta\nu\,, \tag{9–1}$$

where $k_{\rm B} = 1.380 \times 10^{-23}$ Joule K^{-1} is the Boltzmann constant and $\Delta\nu$ is the observing bandwidth.

The power entering the feed is amplified by a factor g^2 where g is the voltage gain (in contrast Crane & Napier [1989] use a power gain G). Thus the powers available from a source and from the system noise are given by:

$$P_{\rm a} = g^2\, k_{\rm B}\, T_{\rm a}\, \Delta\nu \tag{9–2}$$

and

$$P_{\rm N} = g^2\, k_{\rm B}\, T_{\rm sys}\, \Delta\nu\,, \tag{9–3}$$

where $T_{\rm a}$ is the "antenna temperature" for the source and $T_{\rm sys}$ is the "system temperature". $T_{\rm sys}$ includes contributions from the receiver noise, feed losses,

[1] 1 Jansky = 1 Jy = 10^{-26} W m^{-2} Hz^{-1} = 10^{-23} erg sec^{-1} cm^{-2} Hz^{-1}

spillover, atmospheric emission, galactic background and cosmic background (Crane & Napier 1989). The power from the source can be related to the flux density, S, the area of the antenna, A, and the antenna efficiency, $\eta_{\rm a}$, using:

$$P_{\rm a} = \frac{1}{2}\, g^2\, \eta_{\rm a}\, A\, S\, \Delta\nu = g^2\, k_{\rm B}\, K\, S\, \Delta\nu\,, \tag{9–4}$$

where $K = (\eta_{\rm a}\, A) \,/\, (2\, k_{\rm B})$. The factor of 2 in Equation 9–4 accounts for the fact that a single channel receiver is only able to accept half of the total radiation from an unpolarized source. From Equations 9–2 and 9–4, it can be seen that K is simply the gain, or "sensitivity", of the antenna in degrees Kelvin of antenna temperature per Jansky of flux density. The term "sensitivity" is placed in quotes, because the use of the term differs from the description of sensitivity given in Section 1. K is a measure of antenna performance.

It is often instructive to express the system temperature in terms of the system equivalent flux density, $SEFD$, defined as the flux density of a source that would deliver the same amount of power:

$$SEFD = \frac{T_{\rm sys}}{K}\,. \tag{9–5}$$

The $SEFD$ takes into account the efficiency and the collecting area of the antenna, plus the system noise. Thus the $SEFD$ is a useful overall measure of system performance. The $SEFD$ also has the advantage that it can be measured easily by determining the fractional increase in power obtained when going on and off a source of known flux density. Measurements of the system temperature and the gain are each uncertain because they both depend on the system calibration, usually a reference noise source at temperature $T_{\rm c}$. The system calibration cancels out for the $SEFD$. Table 9–1 lists some example $SEFD$s, determined empirically for 19 antennas used at a frequency ν = 5.0 GHz in a VLBI array in 1992 and 1993 (Taylor et al. 1994).

3. What is the Sensitivity of an Interferometer?

3.1. Real and Complex Correlators

Consider the sensitivity of a simple interferometer, with a single real output consisting of the product of the voltages from each of two antennas. For a complex correlator the results of this analysis will apply separately to each output channel. This calculation is commonly made using the autocorrelation function of the product (Crane & Napier 1989). Here an alternate and more intuitive derivation will be used, based on the root-mean-square (RMS) variation of a large number of samples of a Gaussian random variable. This derivation is taken directly from Walker (1995). The reader might also want to consult Wall (1979) for a basic introduction to the Gaussian, or Normal, distribution.

The voltage from antenna i before sampling is the sum of voltages $s_{\rm i}$ from the source and $n_{\rm i}$ from noise. The power from antenna i is given by a constant a, which includes the gain, times the expectation value of the square of the voltage:

$$\langle P_{\rm i}\rangle \;=\; a_{\rm i}\left\langle (s_{\rm i} + n_{\rm i})^2 \right\rangle$$

Table 9–1. Empirical *SEFD*s from Taylor et al. (1994)

Antenna Location	Diameter (m)	*SEFD* (Jy)
NRAL, Cambridge, UK	32	140
NRAL, Jodrell Bank, UK	26	366
MPIfR, Effelsberg, Germany	100	39
OSO, Onsala, Sweden	26	757
NFRA, WSRT, Netherlands	5×25	133
IRA, Medicina, Italy	32	225
IRA, Noto, Italy	32	221
Haystack, Westford, MA, USA	36	606
NRAO, Green Bank, WV, USA	43	126
NRAO, VLA, NM, USA	25	319
NRAO, Saint Croix, VI, USA	25	255
NRAO, Hancock, NH, USA	25	259
NRAO, North Liberty, IA, USA	25	300
NRAO, Fort Davis, TX, USA	25	308
NRAO, Los Alamos, NM, USA	25	270
NRAO, Pie Town, NM, USA	25	280
NRAO, Kitt Peak, AZ, USA	25	308
NRAO, Owens Valley, CA, USA	25	249
NRAO, Brewster, WA, USA	25	281

$$\begin{aligned} &= a_{\rm i}\left[\langle s_{\rm i}^2\rangle + \langle n_{\rm i}^2\rangle\right] \\ &= g_{\rm i}^2\, k_{\rm B}\,(T_{\rm ai} + T_{\rm sysi})\,\Delta\nu \\ &= g_{\rm i}^2\, k_{\rm B}\,(K_{\rm i}S_{\rm T} + T_{\rm sysi})\,\Delta\nu\,, \end{aligned} \tag{9–6}$$

as the voltages from the source and from the noise are not correlated, so the expectation value of the cross term is zero. It is the total flux density seen by the antenna, $S_{\rm T}$, that counts. For simplicity, assume that $S_{\rm T}$ is the same for all antennas, although this may not be true for very large sources that are resolved by the primary beams of the larger antennas.

The power after the cross multiplication in the correlator can be obtained in a similar way:

$$\begin{aligned} \langle P_{\rm ij}\rangle &= \frac{\sqrt{a_{\rm i}a_{\rm j}}}{\eta_{\rm s}}\Big\langle\, (s_{\rm i}+n_{\rm i})(s_{\rm j}+n_{\rm j})\Big\rangle \\ &= \frac{\sqrt{a_{\rm i}a_{\rm j}}}{\eta_{\rm s}}\,\langle s_{\rm i}s_{\rm j}\rangle \\ &= \frac{g_{\rm i}\,g_{\rm j}}{\eta_{\rm s}}\sqrt{K_{\rm i}\,K_{\rm j}}\,k_{\rm B}\,\Delta\nu\,S_{\rm c}\,, \end{aligned} \tag{9–7}$$

since only the signal voltages are correlated. Here the relevant flux density is the correlated flux density, $S_{\rm c}$, which may be less than the total flux density, $S_{\rm T}$. The system efficiency factor, $\eta_{\rm s}$, accounts for various losses in the electronics and digital equipment (Crane & Napier 1989; Walker 1989, 1995).

To obtain the signal-to-noise ratio (SNR), the RMS fluctuations of the correlator output are required. The approach that will be taken here to derive this quantity is to consider the RMS fluctuations of the product of the antenna voltages for each sample. Then, the RMS of the output is presumed to be reduced by the square root of the number of independent samples. A square bandpass of width $\Delta\nu$ and a correlator accumulation time τ_{acc} are assumed so that the number of independent samples, by the Nyquist theorem, is $2\,\Delta\nu\,\tau_{\mathrm{acc}}$. Using the fact that the square of the RMS fluctuation of a Gaussian random variable (in this case, the product from the correlator) is equal to the expectation value of the variable squared minus the square of the mean, one finds:

$$\sigma^2(P_{\mathrm{ij}}) = \frac{a_{\mathrm{i}} a_{\mathrm{j}}}{\eta_{\mathrm{s}}^2}\left\langle \Big[(s_{\mathrm{i}}+n_{\mathrm{i}})(s_{\mathrm{j}}+n_{\mathrm{j}})\Big]^2 \right\rangle - \frac{g_{\mathrm{i}}^2\, g_{\mathrm{j}}^2}{\eta_{\mathrm{s}}^2}\, K_{\mathrm{i}}\, K_{\mathrm{j}}\, k_{\mathrm{B}}^2\, S_{\mathrm{c}}^2\, \Delta\nu^2 \tag{9–8}$$

Using the results above, plus a relation for expanding the expectation value of a product of four variables into combinations of expectation values of products of two variables (Thompson, Moran, & Swensen 1986, p. 156), one obtains:

$$\begin{aligned}
\sigma^2(P_{\mathrm{ij}}) &= \frac{a_{\mathrm{i}} a_{\mathrm{j}}}{\eta_{\mathrm{s}}^2}\left[2\Big\langle(s_{\mathrm{i}}+n_{\mathrm{i}})(s_{\mathrm{j}}+n_{\mathrm{j}})\Big\rangle^2 + \Big\langle(s_{\mathrm{i}}+n_{\mathrm{i}})^2\Big\rangle\Big\langle(s_{\mathrm{j}}+n_{\mathrm{j}})^2\Big\rangle\right] \\
&\quad -\frac{g_{\mathrm{i}}^2\, g_{\mathrm{j}}^2}{\eta_{\mathrm{s}}^2}\, K_{\mathrm{i}}\, K_{\mathrm{j}}\, k_{\mathrm{B}}^2\, \Delta\nu^2\, S_{\mathrm{c}}^2 \\
&= 2\frac{g_{\mathrm{i}}^2 g_{\mathrm{j}}^2}{\eta_{\mathrm{s}}^2}\, K_{\mathrm{i}}\, K_{\mathrm{j}}\, (k_{\mathrm{B}}\, \Delta\nu\, S_{\mathrm{T}})^2 \\
&\quad + \frac{g_{\mathrm{i}}^2\, g_{\mathrm{j}}^2}{\eta_{\mathrm{s}}^2}\, (k_{\mathrm{B}}\, \Delta\nu)^2 (K_{\mathrm{i}}\, S_{\mathrm{T}} + T_{\mathrm{sysi}})(K_{\mathrm{j}}\, S_{\mathrm{T}} + T_{\mathrm{sysj}}) \\
&\quad - \frac{g_{\mathrm{i}}^2\, g_{\mathrm{j}}^2}{\eta_{\mathrm{s}}^2}\, K_{\mathrm{i}}\, K_{\mathrm{j}}\, k_{\mathrm{B}}^2\, \Delta\nu^2\, S_{\mathrm{c}}^2 \qquad\qquad (9\text{–}9) \\
&= k_{\mathrm{B}}^2\, \Delta\nu^2 \frac{g_{\mathrm{i}}^2\, g_{\mathrm{j}}^2}{\eta_{\mathrm{s}}^2} \\
&\quad \times \Big(K_{\mathrm{i}}\, K_{\mathrm{j}}\, S_{\mathrm{c}}^2 + K_{\mathrm{i}}\, K_{\mathrm{j}}\, S_{\mathrm{T}}^2 + K_{\mathrm{i}}\, S_{\mathrm{T}}\, T_{\mathrm{sysi}} + K_{\mathrm{j}}\, S_{\mathrm{T}}\, T_{\mathrm{sysj}} + T_{\mathrm{sysi}}\, T_{\mathrm{sysj}}\Big).
\end{aligned}$$

To write the noise level in units of Janskys, take the square root and divide by (a) the product of $g_{\mathrm{i}}\, g_{\mathrm{j}} \sqrt{K_{\mathrm{i}}\, K_{\mathrm{j}}}\, k_{\mathrm{B}}\, \Delta\nu$, to convert from source flux density to cross correlated power (Equation 9–7); and (b) $\sqrt{2\,\Delta\nu\,\tau_{\mathrm{acc}}}$, the standard deviation of the mean:

$$\Delta S_{\mathrm{ij}} = \frac{1}{\eta_{\mathrm{s}}\sqrt{2\,\Delta\nu\,\tau_{\mathrm{acc}}}} \sqrt{S_{\mathrm{c}}^2 + S_{\mathrm{T}}^2 + S_{\mathrm{T}}\left(\frac{T_{\mathrm{sysi}}}{K_{\mathrm{i}}} + \frac{T_{\mathrm{sysj}}}{K_{\mathrm{j}}}\right) + \frac{T_{\mathrm{sysi}}\, T_{\mathrm{sysj}}}{K_{\mathrm{i}}\, K_{\mathrm{j}}}}\,. \tag{9–10}$$

This is the same result as obtained by Crane & Napier (1989) using the autocorrelation of the cross product.

The above result can be generalized to non-square bandpasses by considering Equation 9–10 to apply to an infinitesimally small bandpass, over which the square bandpass assumption is a good approximation. If $g_{\mathrm{i}}(\nu)$ is the voltage gain as a function of frequency for antenna i, then the overall bandpass used to set the Nyquist sample interval is $\int_0^\infty g_{\mathrm{i}}(\nu)\, g_{\mathrm{j}}(\nu)\, d\nu$. If the gains are carried through the above analysis, $\Delta S_{\rm ij}$ becomes:

$$\Delta S_{\rm ij} = \frac{\sqrt{\displaystyle\int_0^\infty g_{\rm i}^2(\nu)\, g_{\rm j}^2(\nu) \left[S_{\rm c}^2 + S_{\rm T}^2 + S_{\rm T} \left(\tfrac{T_{\rm sysi}}{K_{\rm i}} + \tfrac{T_{\rm sysj}}{K_{\rm j}} \right) + \tfrac{T_{\rm sysi}\, T_{\rm sysj}}{K_{\rm i}\, K_{\rm j}} \right] d\nu}}{\eta_{\rm s} \sqrt{2\, \tau_{\rm acc} \int_0^\infty g_{\rm i}(\nu)\, g_{\rm j}(\nu)\, d\nu}} \qquad (9\text{–}11)$$

Note that the noise of the correlated signal, $S_{\rm c}$, and of the source power as it adds to the total powers at each antenna, $S_{\rm T}$, both contribute to the total noise of the correlated output.

Returning to the assumption of square bandpasses, Equation 9–10 can be applied to two limiting cases: strong sources and weak sources. If the celestial source contributes most of the total noise, then terms involving system temperatures and gains can be ignored. For such cases it is common that $S_{\rm T} \gg S_{\rm c}$, so then Equation 9–10 simplifies to:

$$\Delta S_{\rm ij} = \frac{S_{\rm T}}{\eta_{\rm s} \sqrt{2\, \Delta\nu\, \tau_{\rm acc}}} \, . \qquad (9\text{–}12)$$

In most circumstances of interest, however, the source only contributes a small fraction of the total noise and the terms in Equation 9–10 involving flux density can be ignored:

$$\Delta S_{\rm ij} = \frac{1}{\eta_{\rm s}} \sqrt{\frac{T_{\rm sysi}\, T_{\rm sysj}}{2\, \Delta\nu\, \tau_{\rm acc}\, K_{\rm i}\, K_{\rm j}}} \qquad (9\text{–}13)$$

or in terms of the $SEFD$s defined in Equation 9–5:

$$\Delta S_{\rm ij} = \frac{1}{\eta_{\rm s}} \sqrt{\frac{SEFD_{\rm i}\, SEFD_{\rm j}}{2\, \Delta\nu\, \tau_{\rm acc}}} \qquad (9\text{–}14)$$

For two antennas with the same $SEFD$s:

$$\Delta S_{\rm ij} = \frac{1}{\eta_{\rm s}} \frac{SEFD}{\sqrt{2\, \Delta\nu\, \tau_{\rm acc}}} \qquad (9\text{–}15)$$

For the usual case of a complex correlator, the above analysis applies to each channel individually, with $S_{\rm c}$ corresponding to the appropriate component of the complex visibility. The two output channels are usually called either the sine and cosine channels or the real and imaginary channels. The latter terminology is adopted here, and $S_{\rm R}$ and $S_{\rm i}$ are used to refer to the real and imaginary correlator outputs, respectively, calibrated in Janskys.

As an example, the upper panels of Figure 9–1 show $S_{\rm R}$ and $S_{\rm i}$ from the VLBA correlator at a frequency $\nu = 8.4$ GHz, over the course of about one-half hour on a phase reference source. Does the observed point-to-point scatter match that predicted from Equation 9–15? For these VLBA data $\tau_{\rm acc} = 2$ s and $SEFD = 301$ Jy (Wrobel 1998). Furthermore, the data were recorded in a 2-bit VLBA mode but, since $\eta_{\rm s}$ has not been derived for such data, conversion to the equivalent 1-bit case is made (Walker 1989). This means the effective $\Delta\nu =$ 16 MHz and $1/\eta_{\rm s} = 1.8$. A further small correction of 0.637/0.623 is also made

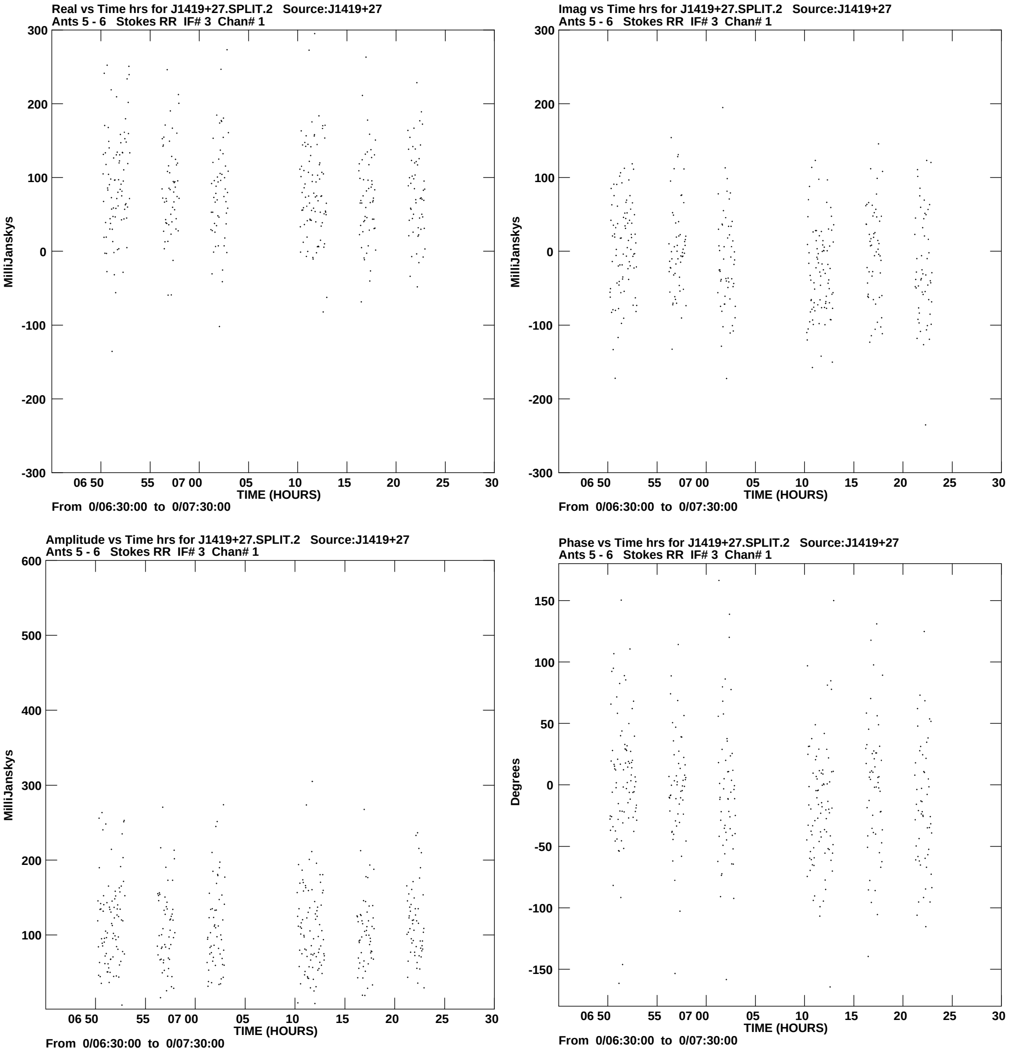


Figure 9–1. Abscissa - Time. Ordinate - *Top left panel:* $S_{\rm R}$ from -300 to 300 mJy. *Top right panel:* $S_{\rm i}$ from -300 to 300 mJy. *Bottom left panel:* $S_{\rm m}$ from 0 to 600 mJy. *Bottom right panel:* $\phi_{\rm m}$ from -180 to 180 degrees.

for the SNR loss of 2-bit data in comparison to 1-bit data (Walker 1989). Then application of Equation 9–15 to this VLBA example yields $\Delta S_{\rm ij}$ = 69 milliJy (mJy), consistent with the point-to-point scatter evident in Figure 9–1 for $S_{\rm R}$ (upper left panel) and for $S_{\rm i}$ (upper right panel). A more quantitative comparison between observed and predicted sensitivities will be made in Section 4, where the image of the target source for this phase-referenced VLBA project will be analyzed.

3.2. Amplitudes and Phases

The complex visibilities are the quantities of interest for imaging. However, for some other applications the measured amplitude, $S_{\rm m}$, and measured phase, $\phi_{\rm m}$, are of interest:

$$S_{\rm m} = \sqrt{S_{\rm R}^2 + S_{\rm i}^2} \quad \text{and} \quad \phi_{\rm m} = \tan^{-1}(S_{\rm i}/S_{\rm R}).$$

An example would be a detection project where one might look for evidence of signal on individual interferometer baselines. Also, some model fitting algorithms deal with the amplitude and phase separately.

The probability distributions for $S_{\rm m}$ and $\phi_{\rm m}$ are (Crane & Napier 1989):

$$P(S_{\rm m}) = \frac{S_{\rm m}}{\Delta S^2} I_0\left(\frac{S_{\rm m}\, S}{\Delta S^2}\right) \exp\frac{-(S_{\rm m}^2 + S^2)}{2\,\Delta S^2} \tag{9–16}$$

and

$$P(\phi - \phi_{\rm m}) = \frac{1}{2\pi}\exp\left(\frac{-S^2}{2\,\Delta S^2}\right)\left(1 + G\sqrt{\pi}e^{G^2}(1 + \mathrm{erf}\, G)\right), \tag{9–17}$$

where S is the true visibility amplitude, ΔS is the standard deviation of the real or imaginary part of the visibility, $G(\theta) = (S\cos\theta)/(\sqrt{2}\,\Delta S)$, I_0 is the modified Bessel function of the first kind and order zero, and erf is the error function. The distribution for the measured amplitude is the same as for a sine wave in noise and is known as a Rice distribution.

If no signal is present then $S = 0$, and Equation 9–16 reduces to a Rayleigh distribution and Equation 9–17 reduces to a uniform distribution. The Rayleigh distribution has a nonzero mean and a standard deviation $\Delta S\sqrt{2 - (\pi/2)}$, while the uniform distribution has a zero mean and a standard deviation of $\pi/\sqrt{3}$ (Thompson, Moran, & Swenson 1986, p. 261). At high SNR, $P(S_{\rm m})$ and $P(\phi - \phi_{\rm m})$ reduce to Gaussian distributions of standard deviation ΔS for the amplitude and $\Delta S/S$ for the phase. At low SNR, the functions are more complicated and, for the amplitude, deliver a Ricean-biased estimate of the true amplitude. Methods to compensate for Ricean bias are available (e. g. Thompson, Moran, & Swenson 1986, p. 262; Killeen, Bicknell, & Ekers 1986). Figure 9–2 shows the probability distributions of the measured amplitude and measured phase for a variety of SNRs. One immediate implication from Figure 9–2 is that weak signals should be much more clearly seen in the phase than in the amplitude. This is demonstrated in the lower panels of Figure 9–1, which show the measured amplitude $S_{\rm m}$ (left panel) and measured phase $\phi_{\rm m}$ (right panel) corresponding to $S_{\rm R}$ and $S_{\rm i}$ plotted in the upper panels: a claim of "detection" is more obvious from the measured phase $\phi_{\rm m}$ than from the measured amplitude $S_{\rm m}$.

4. What is the Sensitivity of a Synthesis Image?

4.1. Derivation

In a phase-referenced observation of a weak target source, the image sensitivity will be the combined sensitivity of all the interferometer combinations integrated over the full time on target. This noise limit will determine the weakest feature that can be detected in the absence of other imaging limitations, such as confusion or dynamic range. The derivation of image sensitivity given below closely follows Crane & Napier (1989) and Walker (1995).

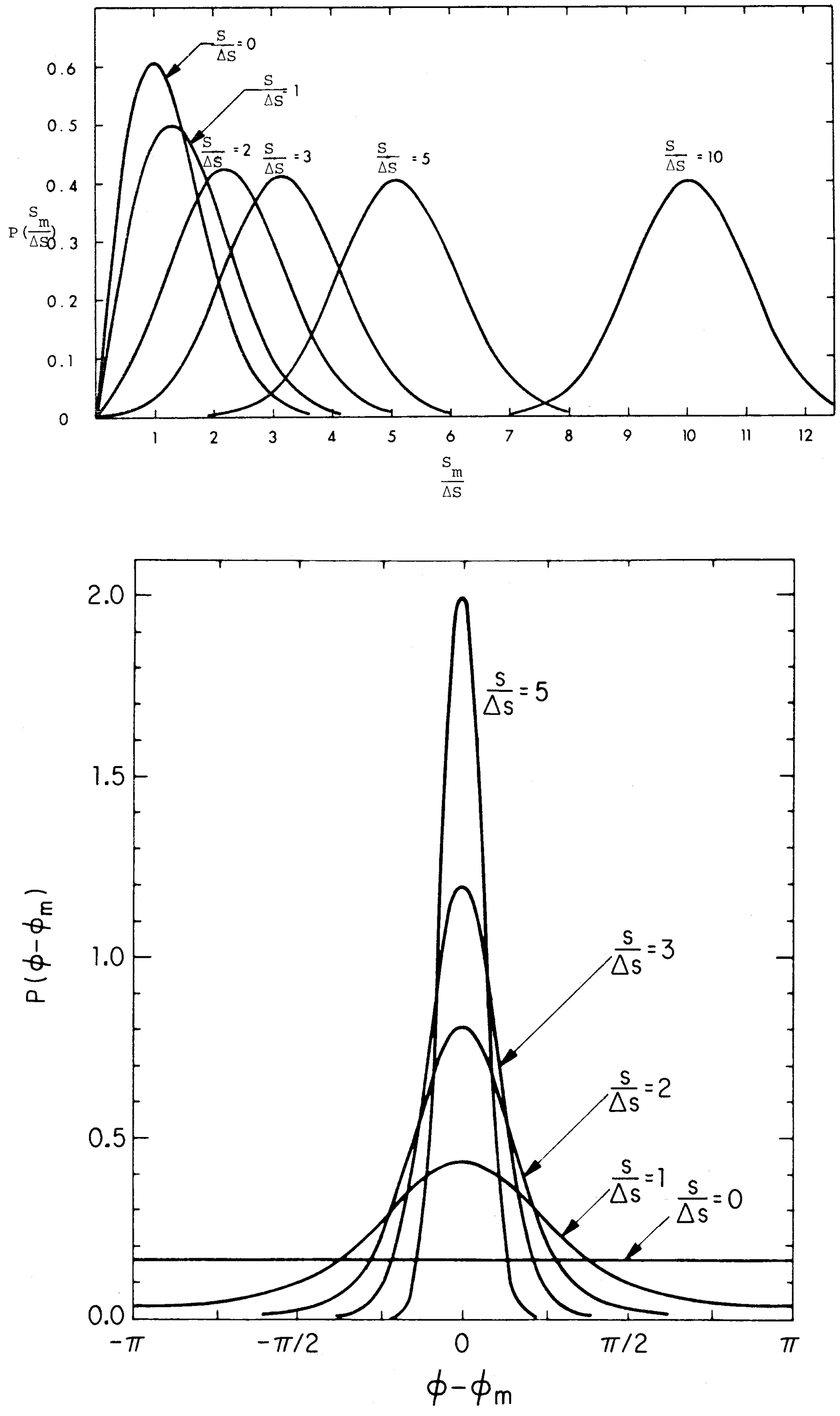


Figure 9–2. *Top panel:* Probability distribution for the amplitudes for a variety of signal-to-noise ratios. *Bottom panel:* Probability distribution for the phases for a variety of signal-to-noise ratios. Both panels are taken from Crane & Napier (1989).

Consider first that each pixel of a single-polarization image is a linear combination of each measured data point:

$$I_{\rm m}(l,m) = C \sum_{\rm k=1}^{2\,L} T_{\rm k}\, W_{\rm k}\, w_{\rm k}\, V_{\rm k}\, e^{2\pi i(u_{\rm k}l+v_{\rm k}m)}. \tag{9–18}$$

Here $I_{\rm m}(l,m)$ is the measured image point, $V_{\rm k}$ is the measured complex visibility data point located at $(u_{\rm k},v_{\rm k})$, C is a normalization constant, $T_{\rm k}$ is the taper function, $W_{\rm k}$ is the weighting function, and $w_{\rm k}$ (not included by Crane & Napier [1989]) is the weight used to reflect the SNR of the data point. The sum is carried out over $2\,L$ points because, since the visibility function is Hermitian, the conjugate points to all measured points are also known. Equation 9–18 has no $k = 0$ term, reflecting an assumption that no zero-spacing flux density is available. It is important to distinguish here between the weighting function $W_{\rm k}$ and the SNR weighting $w_{\rm k}$. The weighting function $W_{\rm k}$ can be unity ("natural weighting"), can reflect the local density in the (u,v) plane of other data points ("uniform weighting"), or can be anywhere in between ("robust weighting"). Weighting by SNR with $w_{\rm k}$ tends to get neglected in discussions of linked interferometers because all antennas are about equally sensitive. For VLBI arrays, however, there can be a very wide range of sensitivities so the $w_{\rm k}$ term can be very important.

To calculate the noise in the image, consider only the central location at $l = 0$ and $m = 0$. In the absence of strong sources, the noise at this point is the same as elsewhere in the image and the math is simplified. At this central pixel, the imaginary part of each data point does not contribute because it is cancelled by the Hermitian conjugate point, so the measured intensity follows from a simple "direct Fourier transform":

$$I_{\rm m}(0,0) = 2\,C \sum_{\rm k=1}^{L} T_{\rm k}\, W_{\rm k}\, w_{\rm k}\, S_{\rm Rk}. \tag{9–19}$$

Each Fourier component contributing to $I_{\rm m}$ is in error because $S_{\rm Rk}$ is in error by $\Delta S_{\rm k}$. This means the associated variance $(\Delta I_{\rm m})^2$ is just the sum of the variances in each Fourier component. Then taking the square route of the sum of the variances gives:

$$\Delta I_{\rm m} = 2\,C \sqrt{\sum_{\rm k=1}^{L} T_{\rm k}^2\, W_{\rm k}^2\, w_{\rm k}^2\, \Delta S_{\rm k}^2}. \tag{9–20}$$

To force the result to come out in terms of flux density per beam area, C is set equal to

$$1\,/\,(2\,\sum_{\rm k=1}^{L} T_{\rm k}\, W_{\rm k}\, w_{\rm k}).$$

For the simplest case of no taper $T_{\rm k} = 1$, natural weighting $W_{\rm k} = 1$, and all points of equal SNR $w_{\rm k} = w$ and equal $\Delta S_{\rm k} = \Delta S$, Equation 9–20 reduces to:

$$\Delta I_{\rm m} \;=\; 2\,C\,\Delta S\, w \sqrt{\sum_{\rm k=1}^{L} 1}$$

$$
\begin{aligned}
&= \frac{2\,\Delta S\, w \sqrt{\sum_{\mathrm{k}=1}^{L} 1}}{2\,w \sum_{\mathrm{k}=1}^{L} 1} \\
&= \Delta S \sqrt{L}/L \\
&= \Delta S/\sqrt{L}, \qquad (9\text{–}21)
\end{aligned}
$$

as expected for the standard deviation of the mean of L samples of a variable with standard deviation ΔS. The number of samples L is just the product of the total number of baselines for an array with N antennas and the total number of correlator accumulation times gathered:

$$L = \frac{1}{2}\, N\, (N-1)\, (t_{\mathrm{int}}/\tau_{\mathrm{acc}}). \qquad (9\text{–}22)$$

Inserting Equations 9–15 and 9–22 into Equation 9–21 then gives:

$$\Delta I_{\mathrm{m}} = \frac{1}{\eta_{\mathrm{s}}} \frac{SEFD}{\sqrt{N\,(N-1)\,\Delta\nu\, t_{\mathrm{int}}}}. \qquad (9\text{–}23)$$

Equation 9–23 is the sensitivity of a single-polarization image formed from a homogeneous array built with N identical antennas. Walker (1989) derives a more general expression for the sensitivity of a single-polarization image formed with an inhomogeneous array of N disparate antennas.

If simultaneous dual polarization data with ΔI_{m} per polarization are available, then the sensitivity of an image of a Stokes parameter I, Q, U, or V will obey Gaussian statistics characterized by a zero mean and a standard deviation:

$$\Delta I = \Delta Q = \Delta U = \Delta V = \frac{\Delta I_{\mathrm{m}}}{\sqrt{2}}. \qquad (9\text{–}24)$$

The noise in an image of the linear polarization flux density

$$P = \sqrt{Q^2 + U^2}$$

will obey Rayleigh statistics, while the noise in an image of the linear polarization position angle

$$\chi = \frac{1}{2} \tan^{-1}(U/Q)$$

will be uniform. These Rayleigh and uniform distributions are discussed by Thompson, Moran, & Swenson (1986, p. 165ff, 261ff) and correspond to the case of no signal ($S = 0$) in Figure 9–2.

4.2. Example: Stokes *I*

Equations 9–23 and 9–24 can be applied to the example Stokes I image shown in the left panel of Figure 9–3. This image, of the Seyfert galaxy NGC 5548, was made with natural weighting and no tapering from VLBA data at a frequency $\nu = 8.4$ GHz. The right panel of Figure 9–3 displays the histogram of image pixel values, excluding pixels contaminated by signal from the Seyfert nucleus. The histogram is a Gaussian distribution with an RMS of 90 microJy beam^{-1}

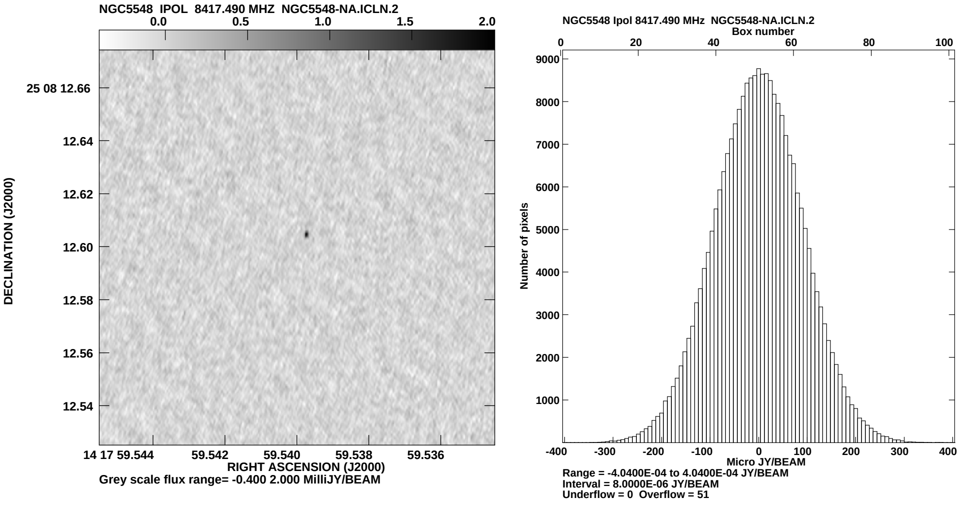


Figure 9–3. *Left panel:* Grey scale display of a VLBA image of Stokes I at a frequency $\nu = 8.4$ GHz of the Seyfert nucleus of NGC 5548 (Wrobel, Conway, & Terlevich, in preparation). Scale range is -400 to $+2000$ μJy beam^{-1}. *Right panel:* Histogram of pixel values spanning 800 μJy beam^{-1} for the image to the left, excluding pixels contaminated by signal from the nucleus.

(μJy beam^{-1}). Does the observed image RMS match that predicted from those equations? The observational parameters are as for the interferometer example discussed in Section 3.1. In addition, NGC 5548 was observed for 22 3-minute scans, with $N = 10$ antennas available for 10 scans, 9 for 9 scans, and 8 for 3 scans. Each antenna accepted an effective $\Delta\nu = 64$ MHz per polarization. Data acquired below an elevation limit of 20 degrees were flagged, so the typical $SEFD = 301$ Jy should be reasonable for all remaining scans. Then application of Equations 9–23 and 9–24 to this VLBA I example predicts $\Delta I = 88$ μJy beam^{-1}, very close to that observed.

Although both image sensitivity and systematic effects will contribute to the error budget for the right ascension and declination of the Seyfert nucleus, what is the positional uncertainty due to the observed image sensitivity alone? The nucleus has a peak brightness $I_{\rm peak} = 2$ mJy beam^{-1} and is unresolved to the VLBA. The apparent brightness distribution of the nucleus therefore mimics the elliptical Gaussian response of the VLBA to a point source. In this example the effective full width of the Gaussian at half maximum, $\theta_{\rm HPBW}$, is about 1.5 millarcsec (mas). The error in the position of a Gaussian peak is $\Delta\theta = (1/2)\,\theta_{\rm HPBW}\,(\Delta I/I_{\rm peak})$ (Ball 1975), so image sensitivity alone limits the positional accuracy to 34 microarcsec.

4.3. Example: Stokes Q and U

Figure 9–4 is meant to illustrate application of Equations 9–23 and 9–24 to some example images of Stokes Q and U, and their relatives P and χ. The upper panels show histograms of pixel values for Stokes Q and U images of the Seyfert galaxy Mrk 231, made with natural weighting and no tapering from VLA data at a frequency $\nu = 1.4$ GHz. Each upper histogram is a Gaussian distribution with an RMS of 17 μJy beam^{-1}. Is this as expected? For these VLA data the typical zenith $T_{\rm sys}$ and $\eta_{\rm a}$ values tabulated by Perley (1997) imply $SEFD$ = 327 Jy. For the VLA correlator in continuum mode $\eta_{\rm s} = 0.79$ (Crane & Napier 1989). Finally, $N = 26$ VLA antennas were available for $t_{\rm int} = 380$ minutes = 22,800 s, and each antenna accepted an effective $\Delta\nu = 23$ MHz (called 25 MHz in VLA lore) per polarization. Then application of Equations 9–23 and 9–24 to this VLA Q and U example predicts $\Delta Q = \Delta U = 16$ μJy beam^{-1}, very close to that observed. The lower panels in Figure 9–4 show histograms of pixel values for the VLA images of P and χ obtained from the appropriate combinations of images of Stokes Q and U. The P histogram in the lower left panel is a Rayleigh distribution, while the χ histogram in the lower right panel is a uniform distribution. These lower panels correspond to the curves plotted in Figure 9–2 for the case of no signal ($S = 0$), except that χ spans half the range of $\phi - \phi_{\rm m}$.

This Mrk 231 example offers a further lesson: the associated Stokes I image has an RMS of 23 μJy beam^{-1} in pixels lacking signal. Equation 9–24 predicts that the images of Stokes I, Q, and U should have the same RMS value if they are limited by sensitivity alone. The higher RMS for Stokes I suggests that additional factors are contributing to the Stokes I noise. Indeed, the ratio of peak-to-RMS observed in the Stokes I image is 10^4, so dynamic range is a likely contributing factor.

4.4. Brightness Temperature Regimes of Arrays

Radio astronomers fixate on temperatures, and it is time to introduce you to yet another one. The "brightness temperature" is defined as the Rayleigh-Jeans temperature, $T_{\rm b}$, of an equivalent black body which will give the same power per unit area per unit frequency interval per unit solid angle as the celestial radio source. The Rayleigh-Jeans relation can be used to derive $T_{\rm b}$ for a source of flux density, S, and solid angle, Ω:

$$T_{\rm b} = \frac{c^2}{2\,k_{\rm B}\,\nu^2}\,\frac{S}{\Omega}, \tag{9–25}$$

where c is the speed of light. The aim is to use this equation to illustrate the brightness temperature regimes of various synthesis arrays, by estimating the *minimum* brightness temperature, $T_{\rm b,min}$, of a source that can be imaged with the array. Let the image sensitivity, ΔI, be taken as the source flux density and the synthesized beam area, $\Omega_{\rm s}$, be set to the *maximum* size of the source responsible for ΔI. If the synthesized beam can be modeled as a circular Gaussian with a full width at half-power, $\theta_{\rm HPBW}$, then

$$\Omega_{\rm s} = \frac{\pi\theta^2_{\rm HPBW}}{4\,{\rm ln}2}, \tag{9–26}$$

and Equation 9–25 can be re-cast as:

$$T_{\rm b,min} = \frac{2\,{\rm ln}2}{\pi}\,\frac{c^2}{k_{\rm B}}\,\frac{\Delta I}{\nu^2\,\theta^2_{\rm HPBW}}, \tag{9–27}$$

Figure 9–5 shows a plot of ΔI as a function of $\theta_{\rm HPBW}$, with sloping lines derived from Equation 9–27 for various $T_{\rm b,min}$ values at a frequency $\nu = 8.4$ GHz. The

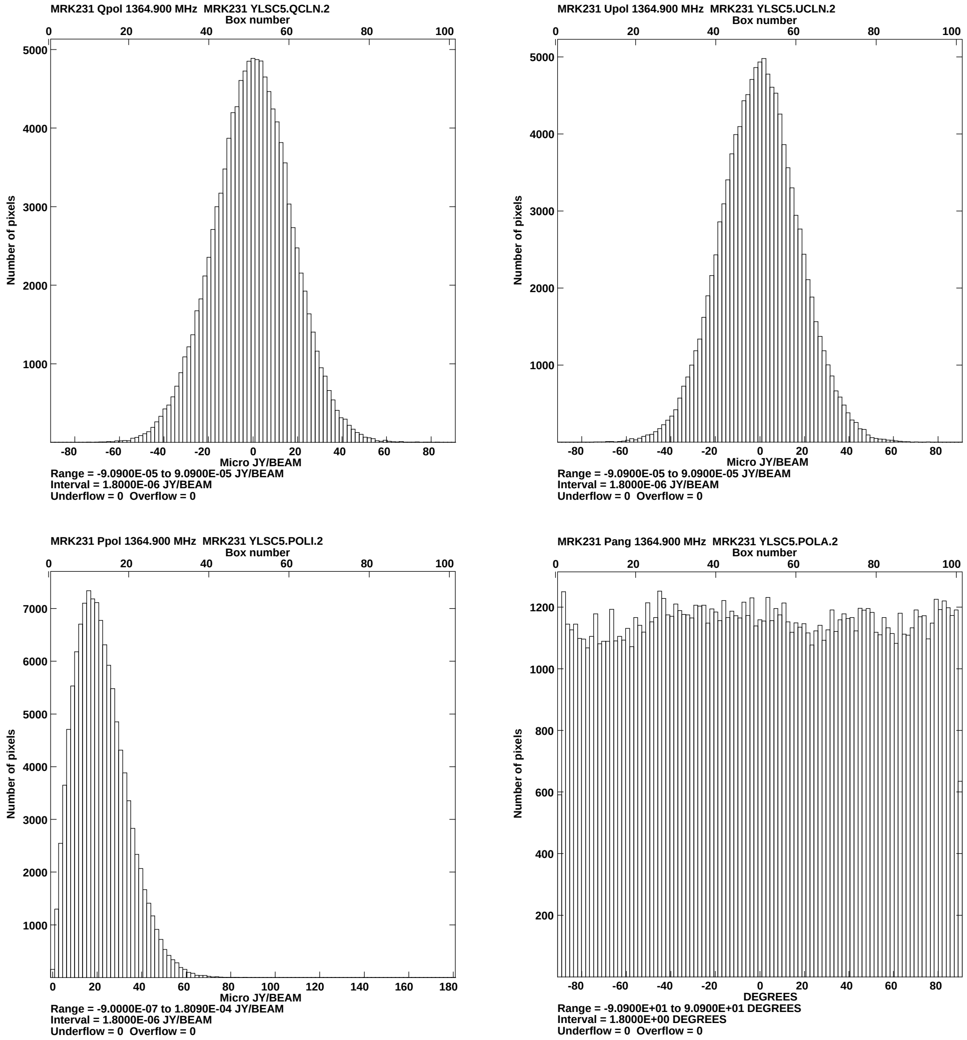


Figure 9–4. Histograms of pixel values with no signal from VLA images at a frequency $\nu = 1.4$ GHz of the Seyfert galaxy Mrk 231 (Ulvestad, Wrobel, & Carilli 1998). *Top left panel:* Stokes Q. Histogram spans 180 μJy beam^{-1} and is centered on the origin. *Top right panel:* Stokes U. Histogram spans 180 μJy beam^{-1} and is centered on the origin. *Bottom left panel:* $P \geq 0$. Histogram spans 180 μJy beam^{-1}. *Bottom right panel:* χ. Histogram spans 180 degrees and is centered on the origin.

plotted letters correspond to the sensitivities and angular resolutions of some synthesis images selected from the recent astronomical literature. This plot demonstrates the brightness temperature regimes of some example arrays.

4.5. Factors Degrading Image Sensitivity

Neglecting imaging limitations such as confusion or dynamic range, what factors can degrade image sensitivity?

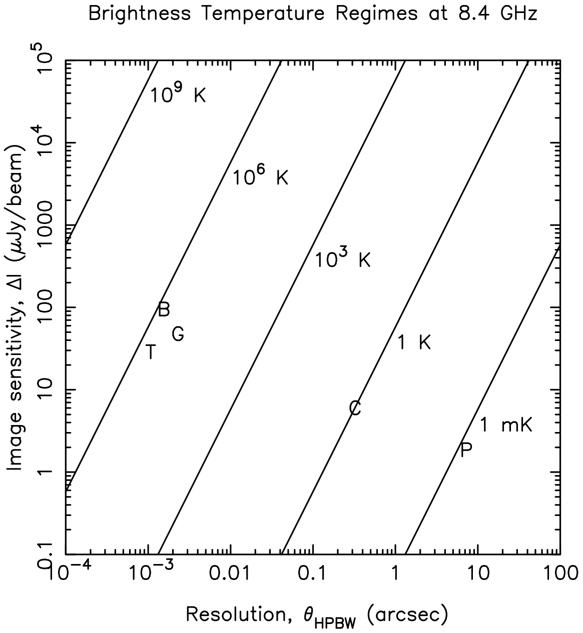


Figure 9–5. Dual-polarization sensitivities, ΔI, as a function of image resolution, $\theta_{\rm HPBW}$, for some published images (C = VLA, Crane et al. 1993; B = VLBA, Blundell et al. 1996; G = VLBA+VLA, Gallimore, Baum, & O'Dea 1997; P = VLA, Partridge et al. 1997; T = VLBA+VLA+Effelsberg, Taylor et al. 1997). The sloping lines correspond to various values for $T_{\rm b,min}$ at a frequency $\nu = 8.4$ GHz.

Some effects can lead to higher noise levels at the edge of an image than at the center of an image (Crane & Napier 1989). The derivation given above for the image center assumed that the image was made from a direct Fourier transform of the visibility data. In practice a fast Fourier transform is usually used, and the associated convolution and gridding in the (u, v) plane will lead to higher noise levels at the image edge. Also, each antenna in an array has its own primary beam gain pattern, causing reduced sensitivity at off-center image locations.

Image sensitivity can also be degraded by the effects of fringe fitting and/or self-calibration (Walker 1995). Errors in the determination of the antenna calibration parameters will introduce errors in the visibility data. For the special case of a relatively strong, nearly unresolved source and N identical antennas, Cornwell & Fomalont (1989) showed that image sensitivity is worse by a factor of $\sqrt{(N-1)/(N-3)}$ when amplitude and phase self-calibration is used than when it is not used. This effect is most pronounced for images formed from visibility data with the same handedness (e. g. RCP, LCP, Stokes I, and Stokes V). For complicated sources, the corresponding analysis cannot be done, but experience indicates that the sensitivity is rarely reduced by a factor of 2 or 3.

Although the lowest noise per pixel can be obtained with natural weighting and no taper, it is often desirable to use other imaging parameters. If the source is resolved, the longest baselines have little or no correlated flux density so using them in the imaging process only adds their noise, but no information about the signal. Therefore, application of a taper such that only those baselines with signal contribute can enhance the detectability of the source. Effectively the resolution is lowered until the source is concentrated into a small number of resolution elements, raising the flux density per resolution element faster than the noise. The prices paid are higher noise per pixel and lower resolution. If too strong a taper is applied, data points with good signal are excluded from the image and the sensitivity is reduced. As an example, Crane & Napier (1989) derive the Gaussian taper function which maximizes the sensitivity of each VLA configuration to a model source with a Gaussian surface brightness. Weighting functions other than natural are also used to affect the quality of the beam. With natural weighting and typical array geometries, most of the data tend to be concentrated toward the center of the (u, v) plane and the resolution is much less than that of the longest baselines. Robust or uniform weighting can be used to reduce the emphasis on short baselines and to give higher resolution. Such weighting schemes also have the effect of smoothing out uneven distributions of points and giving a more Gaussian synthesized beam. The price paid is a non-optimal weighting of the data points, resulting in poorer sensitivity. The actual magnitudes of all these weighting effects depend on the details of the sensitivities of the antennas in the array and on the distribution of points in the (u, v) plane.

5. Summary

- The overall antenna performance is measured by the system equivalent flux density $SEFD$ in units of Janskys.

- Connect two antennas with the same $SEFD$, observing the same bandwidth $\Delta\nu$, into an interferometer with a complex correlator. Let the system efficiency be $\eta_{\rm s}$ and the accumulation time be $\tau_{\rm acc}$. Then the sensitivity of either the real or the imaginary correlator output, in the weak-source limit, is

$$\Delta S_{\rm ij} = \frac{1}{\eta_{\rm s}} \frac{SEFD}{\sqrt{2\,\Delta\nu\,\tau_{\rm acc}}}$$

  in units of Janskys.

- Connect N antennas with the same $SEFD$, observing the same bandwidth $\Delta\nu$, into an array with a complex correlator. Let the system efficiency be $\eta_{\rm s}$. If the array integrates for a time $t_{\rm int}$, then in the weak-source limit the sensitivity of a synthesis image of a single polarization is

$$\Delta I_{\rm m} = \frac{1}{\eta_{\rm s}} \frac{SEFD}{\sqrt{N\,(N-1)\,\Delta\nu\,t_{\rm int}}}$$

  in units of Janskys per synthesized beam area.

Acknowledgments. The National Radio Astronomy Observatory is a facility of the National Science Foundation, operated under a cooperative agreement by Associated Universities, Inc.

References

Ball, J. A. 1975, in *Methods in Computational Physics*, Volume 14, eds. B. Alder, S. Fernbach, & M. Rotenberg (New York: Academic Press), 177–219.

Blundell, K. M., Beasley, A. J., Lacy, M., & Garrington, S. T. 1996, *ApJ*, 468, L91–L94.

Cornwell, T. J., & Fomalont, E. B. 1989, in *Synthesis Imaging in Radio Astronomy* eds. R. A. Perley, F. R. Schwab, & A. H. Bridle (San Francisco: ASP), 185–197.

Crane, P. C., & Napier, P. J. 1989, in *Synthesis Imaging in Radio Astronomy* eds. R. A. Perley, F. R. Schwab, & A. H. Bridle (San Francisco: ASP), 139–165.

Crane, P. C., Cowan, J. J., Dickel, J. R., & Roberts, D. A. 1993, *ApJ*, 417, L61–L62.

Gallimore, J. F., Baum, S. A., & O'Dea, C. P. 1997, *Nature*, 388, 852–854.

Killeen, N. E. B., Bicknell, G. V., & Ekers, R. D. 1986, *ApJ*, 302, 306–336.

Partridge, R. B., Richards, E. A., Fomalont, E. B., Kellermann, K. I., & Windhorst, R. A. 1997, *ApJ*, 483, 38–50.

Perley, R. A. 1997, "The VLA Observational Status Summary", NRAO document.

Taylor, G. B., Vermeulen, R. C., Pearson, T. J., Readhead, A. C. S., Henstock, D. R., Browne, I. W. A., & Wilkinson, P. N. 1994, *ApJS*, 95, 345–369.

Taylor, G. B., Frail, D. A., Beasley, A. J., & Kulkarni, S. R. 1997, *Nature*, 389, 263–265.

Ulvestad, J.S., Wrobel, J.M., & Carilli, C.L. 1998, *ApJ*, in press

Thompson, A. R., Moran, J. M., & Swenson, G. W., Jr. 1986, *Interferometry and Synthesis in Radio Astronomy* (New York: John Wiley & Sons).

Walker, R. C. 1989, in *Very Long Baseline Interferometry* eds. M. Felli & R. E. Spencer (Dordrecht: Kluwer), 163–182.

Walker, R. C. 1995, in *Very Long Baseline Interferometry and the VLBA* eds. J. A. Zensus, P. J. Diamond, & P. J. Napier (San Francisco: ASP), 133–157.

Wall, J. V. 1979, *QJRAS*, 20, 138–152.

Wrobel, J. M. 1998, "The VLBA Observational Status Summary", NRAO document.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

10. Self-Calibration

Tim Cornwell
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Ed B. Fomalont
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A.

Abstract.
In this lecture the principles, techniques, and foibles of self-calibration are discussed.

1. Problems with Ordinary Calibration

Calibrating a synthesis array is one of the most difficult aspects of its operation and, in many cases, is the most important factor in determining the quality of the final deconvolved image. Small quasi-random errors in the amplitude and phase calibration of the visibility data scatter power and so produce an increased level of "rumble" in the weaker regions of the image, and other systematic errors can lead to a variety of artifacts in the image.

The ordinary calibration procedure (see Lecture 5) relies on frequent observations of radio sources of known structure, strength and position in order to determine empirical corrections for time-variable instrumental and environmental factors that cannot be measured, or monitored, directly. The relationship between the visibility $\widetilde{V}_{ij}$ observed at time t on the i-j baseline and the true visibility $V_{ij}(t)$ can be written very generally as

$$\widetilde{V}_{ij}(t) = g_i(t)g_j^*(t)G_{ij}(t)V_{ij}(t) + \varepsilon_{ij}(t) + \epsilon_{ij}(t)\,. \qquad (10\text{–}1)$$

The multiplicative factors $g_i(t)$ and $g_j(t)$ represent the effects of the complex gains of the array elements i and j; $G_{ij}(t)$ is the non-factorable part of the gain on the i-j baseline; $\varepsilon_{ij}(t)$ is an additive offset term; and $\epsilon_{ij}(t)$ is a pure, zero-mean, noise term representing the thermal noise. The effects of $G_{ij}(t)$ and $\varepsilon_{ij}(t)$, which cannot be split into antenna-dependent parts, can usually be reduced to a satisfactory degree by clever design (see Lecture 4), so we will ignore them during this lecture. Equation 10–1 then simplifies to

$$\widetilde{V}_{ij}(t) = g_i(t)g_j^*(t)V_{ij}(t) + \epsilon_{ij}(t)\,. \qquad (10\text{–}2)$$

For simplicity we neglect the effects of time averaging and finite bandwidth, discussed in Lectures 2, 17 and 18; these have relatively little impact here. The *element gain* (usually called the *antenna gain* in radio astronomy) really describes the properties of the elements relative to some reference (usually one array element for phase and a "mean" array element for amplitude). Although this use of the word "gain" may seem confusing, it is quite helpful in lumping all element-based properties together. The gain for any one array element has two contributing components: firstly, a slowly varying instrumental part and secondly, a more rapidly varying part due to the atmosphere (troposphere and ionosphere) above the element. Variations in the phase part of the atmospheric component nearly always dominate the overall variation of the element gains (see Sec. 4 of Lecture 5).

Given a calibration source near the region to be imaged, one can solve for the element gains as functions of time. Interpolation of the solutions then provides approximate values for use in correcting the source visibility data. If the equations are overdetermined, then a least-squares technique can be utilized to good effect in overcoming the random errors embodied in the $\epsilon_{ij}(t)$. In particular, for an array in which data from all baselines are correlated and whose elements are identical, when calibrating on a point source of flux density S the variance in the gain estimates due to the receiver noise is

$$\sigma_G^2 = \frac{\sigma_V^2}{(N-3)S^2}\,, \tag{10–3}$$

where σ_V^2 denotes the variance of a visibility datum (assuming all visibilities have equal variance) and N is the number of array elements (Cornwell 1981).

The main drawback to ordinary calibration arises from temporal and spatial variations in the atmosphere (troposphere and ionosphere) through which the wavefront passes before reaching the array elements. Values for the $g_i(t)$ inferred from observations of a calibration source may not apply to a source observed at a different time and in a different part of the sky. Hence, the effect of the $g_i(t)$'s cannot be removed completely, and residual errors remain. The level of error varies tremendously with the frequency at which the observations are made and with the lengths of the baselines involved, but on a source of appreciable strength it nearly always overwhelms the error due to the receiver noise term.

Other obstacles to ordinary calibration are the strength (or lack of it) of the calibrators, and any resolved structure they may contain. In some circumstances one may not be able to find a sufficiently strong unresolved calibration source anywhere near the source of interest.

The net effect of this calibration problem depends upon the context. In VLBI, it prevents imaging altogether, whereas for shorter-baseline arrays (such as the VLA and Westerbork) it merely lowers the image quality attainable. Fortunately, progress can be made if the element gains are allowed to be degrees of freedom when determining the sky intensity distribution. *Allowing the element gains to be free parameters is the basic principle of self-calibration.*

2. Redundant Calibration and Self-Calibration

We now discuss the pros and cons of letting the element gains be free parameters. If all baselines are correlated then there are, at any one time, N complex gain errors corrupting the $\frac{1}{2}N(N-1)$ complex visibility measurements. Hence there must be at least $\frac{1}{2}N(N-1) - N$ "good" complex numbers hidden in the data that can be used to constrain the true sky intensity distribution. [1] Let us briefly consider what is lost by using only these "good" numbers. The most obvious

[1] Actually, because absolute phase is meaningless for an interferometer, there are $\frac{1}{2}N(N-1) - (N-1)$ "good" phases and $\frac{1}{2}N(N-1) - N$ "good" amplitudes.

losses are the absolute position and strength of the source. The former produces a phase term in the visibility which depends upon the difference in position of the element in an interferometer (see Lecture 1); hence it can be factored out as two element-related quantities. The loss of absolute source strength information is obvious from Equation 10–2. One also loses the ability to distinguish between various different source structures, but we will show that for large enough numbers of array elements this effect is not too important, since the ratio of the number of constraints to number of degrees of freedom increases.

It is clear what one can expect to lose by letting the element gains be free variables, but the many degrees of freedom embodied in the element gains $g_i(t)$ must still be balanced somehow. There are two different schemes: the explicit use of *redundancy*, and the use of *a priori knowledge* about the object. We shall examine these in turn.

2.1. Redundant calibration

Suppose that the geometry of the interferometer array is designed so that some different pairs of array elements measure the same spacing, or (u, v) sample. As an example, consider a one dimensional linear array of N elements equally spaced, with separation d. All spacings except the longest are measured more than once. In fact there are only $N-1$ different spacings measurable, while there are $\frac{1}{2}N(N-1)$ pairs of elements. This redundancy enables solution for both the $N-1$ true visibility samples, up to a linear phase slope, and the N complex gains, again up to a linear phase slope (Hamaker *et al.* 1977). Since the system of equations is overdetermined, a least-squares method can be employed to good effect in suppressing the effects of receiver noise.

Complete redundancy is not necessary for this approach to work; in fact, since only N complex gains need be solved for, there need be only N redundant spacings. The drawback is that the signal-to-noise ratio of the estimated true visibilities decreases, and nulls can prove disastrous.

Redundant calibration is currently used at the Westerbork Synthesis Radio Telescope.

2.2. Self-calibration

The basis of this approach is that in many cases, even after adding the degrees of freedom in the element gains, the estimation of an adequate model of the brightness is still overdetermined (see Lecture 8). Hence self-calibration is really just another method like 'CLEAN' (Lecture 8) which is used to interpret the visibility data by introducing some plausible assumptions about the source structure.

Our aim is to produce a model $\widehat{I}$ of the sky intensity distribution, the Fourier transform $\widehat{V}$ of which, when corrected by some complex gain factors, reproduces the observed visibilities to within the noise level. The model $\widehat{I}$ should be astronomically plausible: for example, possible constraints are positivity of brightness and confinement of the structure. (Other, more elaborate, constraints could involve the maximization of some measures of "goodness" of an image; see Lecture 8). One convenient method (Schwab 1980b) of obtaining such agreement is to minimize—by adjusting both the complex element gains g_i and g_j and the model intensity distribution $\widehat{I}$—the sum of squares of residuals

$$\mathcal{S} = \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) \left| \widetilde{V}_{ij}(t_k) - g_i(t_k) g_j^*(t_k) \widehat{V}_{ij}(t_k) \right|^2 , \tag{10–4}$$

where the $w_{ij}(t_k)$ are weights (purely from signal-to-noise considerations each should be set to the reciprocal of the variance of $\epsilon_{ij}(t_k)$). The time over which the gains should be held constant depends upon the signal-to-noise ratio and upon the variability of the atmosphere (see Sec. 5.3).

An interesting and illuminating connection to ordinary calibration is apparent if Equation 10–4 is re-expressed as

$$\mathcal{S} = \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) |\widehat{V}_{ij}(t_k)|^2 \left| X_{ij}(t_k) - g_i(t_k) g_j^*(t_k) \right|^2 , \tag{10–5}$$

where

$$X_{ij}(t) = \frac{\widetilde{V}_{ij}(t)}{\widehat{V}_{ij}(t)} . \tag{10–6}$$

Division by the model visibilities $\widehat{V}_{ij}(t)$ in effect transforms the object being imaged into a pseudo-point source, though admittedly one with rather strange receiver noise, that can then be used in the ordinary calibration outlined in Section 1.

It is crucial to this gain-solution step that there be too few degrees of freedom (i.e., the element gains $g_i(t)$) to allow the model $\widehat{V}_{ij}(t)$ to be reproduced exactly. If there were, nothing would be achieved. The overdeterminacy also means that errors in the model are averaged down, to an extent dependent on the number of elements in the array. This suggests a possible line of attack in which the model is iteratively refined:

1. Make an initial model of the source using whatever constraints you have on the source structure.

2. Convert the source into a point source using the model.

3. Solve for the complex gains.

4. Find the corrected visibility,

$$V_{ij,\mathrm{corr}}(t) = \frac{\widetilde{V}_{ij}(t)}{g_i(t) g_j^*(t)} . \tag{10–7}$$

5. Form a new model from the *corrected* data, again using constraints upon the source structure.

6. Go to (2), unless you are satisfied with the current model.

This approach divides the optimization problem into a part dealing only with the (u,v) data and a part dealing only with the model of the sky brightness. The former can be solved by a simple iterative approach (Schwab 1980b), and in Lecture 8 we learned that both 'CLEAN' (Sec. 2) and the Maximum Entropy Method (MEM, Sec. 4) can be used to solve the latter problem.

Another view of this iterative approach arises from the application of an optimization approach, such as MEM, to gain correction. The unknown gains are added as free variables in the optimization. In the specific case of MEM, the problem is then to choose the image I_k and the gains $g_i(t)$ so as to maximize the image entropy

$$\mathcal{H} = -\sum_k I_k \ln \frac{I_k}{M_k e}\,, \tag{10–8}$$

subject to

$$\begin{aligned} \mathcal{S} &= \sum_k \sum_{\substack{i,j\\ i\neq j}} w_{ij}(t_k) \left| \widetilde{V}_{ij}(t_k) - g_i(t_k) g_j^*(t_k) \widehat{V}_{ij}(t_k) \right|^2 \\ &= \text{expected value}\,, \end{aligned} \tag{10–9}$$

and

$$\sum_k I_k = \text{estimated value of total flux density}\,, \tag{10–10}$$

where $\widehat{V}_{ij}(t)$ is given by the inverse Fourier transform of the MEM image I_k.

The most general approach to solving this optimization problem would vary the image and the gains simultaneously, whereas the iterative approach consists of alternately fixing either the image or the gains, and varying the other as required. The latter is certainly easier to code and seems to work most of the time.

2.3. Redundant calibration or self-calibration?

The relative merits of redundant calibration and of self-calibration are still being debated. The real question is not: *Should redundant calibration be used with an existing array?* (of course it should, if that is possible), but rather: *Should new arrays be designed with redundant spacings?* The main advantage of redundant calibration is that the results are almost model-independent (there is a variable phase shift to worry about), but it is less flexible than self-calibration, and it uses the available signal-to-noise ratio rather less efficiently. A compromise would be to use redundant calibration to get the structure basically correct, and then to use self-calibration to improve the signal-to-noise. In practice, self-calibration is more commonly used simply because many arrays are not instantaneously redundant. Therefore in the rest of this lecture we will concentrate on self-calibration. First, however, we digress slightly to emphasize the links of both schemes with other methods of phase correction.

3. Other Approaches to Phase Correction

The two schemes for phase correction described in Section 2 have two close relatives: the concept of *closure*, and *adaptive optics*.

3.1. Closure quantities

In the early days of radio interferometry, Roger Jennison was faced with the problem of measuring phase information with interferometers which were inherently phase-unstable. He was struck by the fact that an appropriate sum of visibility phases around a closed loop of baselines is free of element-related errors (Jennison 1953, 1958). This can be confirmed by taking the phase part of Equation 10–2,

$$\widetilde{\phi}_{ij}(t) = \phi_{ij}(t) + \theta_i(t) - \theta_j(t) + \text{noise term}\,, \tag{10–11}$$

where $\theta_i(t) = \arg g_i(t)$. Now suppose that a loop of three baselines is formed from elements i, j and k. Then the quantity $\widetilde{C}_{ijk}(t)$, known as the observed *closure phase* [2], is given by

$$\begin{aligned}\widetilde{C}_{ijk}(t) &= \widetilde{\phi}_{ij}(t) + \widetilde{\phi}_{jk}(t) + \widetilde{\phi}_{ki}(t) \\ &= \phi_{ij}(t) + \phi_{jk}(t) + \phi_{ki}(t) + \text{noise term} \\ &= C_{ijk}(t) + \text{noise term}\,.\end{aligned} \tag{10–12}$$

Thus, for an array of three or more elements, and neglecting noise, closure phase is always a good observable. For an array composed of N elements there are $\frac{1}{2}N(N-1) - (N-1)$ independent closure phases; these are just the "good" constraints mentioned in Section 2.

A *closure amplitude* Γ_{ijkl} can be defined for any loop of 4 elements:

$$\Gamma_{ijkl}(t) = \frac{|\widetilde{V}_{ij}(t)|\,|\widetilde{V}_{kl}(t)|}{|\widetilde{V}_{ik}(t)|\,|\widetilde{V}_{jl}(t)|}\,. \tag{10–13}$$

The amplitudes of the complex gains cancel out of these ratios. Thus, apart from noise, the observed and true closure amplitudes should be identical. There are $\frac{1}{2}N(N-1) - N$ such closure amplitudes.

These closure quantities were of little use until the advent of sufficiently fast computers. Neither closure quantity can be used directly to form an image. However, in the 1970s iterative schemes were developed by Readhead & Wilkinson (1978), Cotton (1979) and others to produce 'CLEAN' images consistent with the closure quantities—see Ekers (1984) for an account of the history of closure phase and self-calibration.

Readhead and Wilkinson (RW) used the following approach to incorporate the closure phases:

1. Make an initial model of the source.

2. For all independent closure phases, use the model to provide estimates of the true phases on two baselines and derive the phase on the other baseline in the loop from the observed closure phase.

[2] This terminology is similar to that of closing, or closure, errors in traversed loops, used by surveyors.

3. Form a new model, using 'CLEAN', from the observed visibility amplitudes and the predicted visibility phases.

4. Go to (2), unless you are satisfied with the current model.

Readhead *et al.* (1980) developed a similar algorithm to include the closure amplitudes as constraints. The aspect of choice in part (2) was eliminated in Cotton's (1979) algorithm by utilizing a least-squares technique.

These various approaches have been widely used in VLBI to produce so-called *hybrid images* from the poorly calibrated data that are commonly collected. Only three serious drawbacks are present in the RW–Cotton type algorithms:

1. Proper treatment of noise is difficult because it occurs additively in the vector visibility, not in the amplitude or phase (see Eq. 10–2). Thus it obeys a simple normal distribution in the vector, but a much more-complicated Rice distribution in the phase.

2. For any array with a large number of elements there are very many more possible than independent closure quantities. For a source showing significant structure, the different closure quantities will have varying signal-to-noises, and so in the RW approach it is not easy to choose an optimal set of closure quantities.

3. Calibration effects in radio imaging really do occur in relation to antennas, not baselines, so incorporation of other constraints on, for example, the variability of the atmospheric phase, is simplest in an element-based approach (Cornwell & Wilkinson 1981).

All of these disadvantages are overcome in self-calibration which, since it alters only element gains, must conserve the closure quantities and thus is equivalent to the use of closure quantities (Cornwell & Wilkinson 1981)

3.2. Adaptive optics

Optical "antennas" are typically limited to about one-arcsecond resolution by rapidly varying path length fluctuations due in turn to variations in the refractive index of air (see Woolf 1982 for a good description). One recently developed technique for overcoming this distortion is known as adaptive optics; a well-chosen name since the optics of the element are distorted in order to cancel the effects of the path length variations. A "rubber mirror", which can be distorted at rates up to 1 kHz, is inserted into the light path, and its shape is controlled by a feedback loop designed to optimize the quality of the final image (see, e.g., Muller & Buffington 1974). One of the measures of quality is the sharpness, defined to be the sum of the squares of the pixel values. In an interesting paper, Hamaker *et al.* (1977) show that in redundant spacing interferometry the sharpness is maximized by requiring that all redundant spacings yield the same visibility phase, exactly the same requirement as used in Section 2.1.

The connection between adaptive optics and the scheme outlined in Section 2.2 should be obvious. In both approaches, the phase of the array element is seen as a free variable which can be changed to obtain a plausible image.

Fortunately, at radio wavelengths the "fringes" (complex visibilities) can be recorded for each interferometer and the correction can be made subsequently, rather than in real time. Furthermore, since "fringes", rather than the image, can be recorded we can keep track of which pair of elements produced each datum. Dyson (1975) has investigated the latter point in relation to adaptive optics; he has shown that interferometer-based correction requires only one photon per atmospheric coherence time per aperture patch to be corrected, while the image-based correction scheme requires the same rate *per pair* of patches. In the latter the extra photons are lost to decorrelation.

4. Why Does Self-Calibration Work?

No proof of convergence has ever been given for self-calibration, so the exact circumstances under which it works are unknown. Such a proof would be very difficult because of the required use of nonlinear methods of deconvolution such as 'CLEAN' to enforce constraints on the source structure. We do however understand *qualitatively* why it works. There are two, related, reasons:

1. Self-calibration is most successful for arrays with large numbers of elements. The ratio of visibility constraints to unknown gains, $\frac{N-2}{2}$ for phases and $\frac{N(N-3)}{2(N-1)}$ for amplitudes, rises without bound as N increases. Consequently, by allowing the calibration to be a variable only a small amount of information is lost.

2. Sources are relatively simple and can be well represented by a small number of degrees of freedom (in the case of 'CLEAN', the parameters specifying the 'CLEAN' components). Hence the source is, in many cases, effectively oversampled and we can afford to introduce a small number of extra degrees of freedom (the antenna gains). The other side of this is that the (u, v) coverage is usually quite good for the simple sources we are interested in.

The basic requirement is that the total number of degrees of freedom (the number of free gains plus the number of free parameters in the model of the sky brightness distribution), should not be greater than the number of independent visibility measurements (see Lecture 8 for further details).

Self-calibration fails when either the signal-to-noise ratio is too low or the source is too complex (relative to the model). Quantitative estimates of the signal-to-noise requirements can be made; whereas the effect of source complexity is much more difficult to estimate, and further work is needed in this area.

5. Practical Problems in Self-Calibration

We will now consider the details of controlling the self-calibration process. Of all the steps involved in image construction, self-calibration is probably the easiest to perform incorrectly, and so a certain amount of care must be employed when choosing the various parameters. Many of these steps are also described, in more detail, in Lecture 13.

5.1. Specifying the model

In the early days of hybrid imaging great care was taken when producing, usually by model-fitting to the amplitudes, an initial model of the sky brightness; the subsequent convergence depended strongly upon the quality of this model. However, experience with self-calibration algorithms used on data from arrays with relatively modest numbers of elements, such as MERLIN, indicates that for a reasonably simple source, use of an initial point source model may delay but will not prevent convergence—see Cornwell and Wilkinson (1981), for example.

Partially phase-stable arrays such as the VLA usually produce visibility data which, on initial imaging and 'CLEAN'ing, give 'CLEAN' component models which can be used to start self-calibration (even though the associated 'CLEAN' images have only modest dynamic range—typically 10–20 dB).

At any stage in self-calibration *it is important to exclude any features of the model that are due to the very calibration errors we wish to eliminate.* Otherwise, the calibration errors will just be passed through from one iteration to the next. A good rule of thumb when constructing a model from 'CLEAN' components is to exclude all components found after the first negative one. [3] The same rule usually works well in subsequent passes through the self-calibration process. Thus the role that 'CLEAN' or MEM plays in rejecting unsatisfactory models of the sky brightness is apparent; if one used a deconvolution method which did not at least partially reject artifacts due to calibration errors, self-calibration could not increase the dynamic range.

Since the model does not have to be very accurate, an image taken at another frequency will often be useful in speeding convergence. Also, for arrays with many elements, a model made at a higher resolution may be adequate.

5.2. Type of solution and weighting schemes

One can sometimes help convergence by choosing whether to solve Equation 10–4 only for the phases or for both amplitudes and phases. Different weighting schemes can be used to emphasize different parts of the model.

Initially, although the phase errors are usually dominant, the model may represent the true visibility phases very well but the amplitudes very poorly. One such example is the use of a point source model for a symmetrical source such as a Gaussian. Correction of the amplitudes using such a model could produce severe errors in subsequent models. Experience shows that in most cases the quality of the fit of a model to the amplitudes is inferior to the fit to the phases, and so it is often prudent to solve initially for the phase errors only.

The form of the weights can be used to control the solution: in the preferred "natural" weighting scheme, the weights $w_{ij}(t)$ in Equation 10–4 are set to the reciprocal of the expected variance of the errors. The effect of weak visibility points is thus decreased; for visibility functions containing nulls this can be important. If the model has systematic errors then it may be advantageous to make the weights depend upon the (u, v) coordinates. For example, suppose that at high resolution the source is well represented but that an additional amount of extended emission is present. By setting $w_{ij}(t)$ to zero for $\sqrt{u^2 + v^2}$ less than

[3] See Lecture 13 for discussion of possible exceptions to this rule.

some limit dependent on the source structure we may obtain better estimates for the gain errors than those which would be obtained from all the data.

5.3. Self-calibration averaging time

Either $\widetilde{V}_{ij}(t)$ or $X_{ij}(t)$ can be averaged over a finite time interval to improve the signal-to-noise ratio. Note that averaging of $X_{ij}(t)$ will not, in general, produce the best signal-to-noise ratio but will correct phase winding that is due to position errors or offsets.

The choice of the optimal averaging time τ_{sc} obviously depends upon the timescale for gain changes and upon the source strength. The error in the gain estimate due to the receiver noise on a nearly unresolved source is (for good signal-to-noise ratio), for amplitude and phase correction,

$$\sigma_G^2(\tau_{\mathrm{sc}}) = \frac{\sigma_V^2(\tau_{\mathrm{sc}})}{(N-3)S^2}\,, \tag{10–14}$$

and, for phase correction,

$$\sigma_G^2(\tau_{\mathrm{sc}}) = \frac{\sigma_V^2(\tau_{\mathrm{sc}})}{(N-2)S^2}\,, \tag{10–15}$$

where S is the approximate flux density of the source, and $\sigma_V^2(\tau)$ is the variance of the receiver noise on each baseline as a function of integration time τ (see Cornwell 1981 for the derivation). One interpretation is that the r.m.s. error in the calculation of the gain of an antenna is approximately the reciprocal of the signal-to-noise ratio for each antenna.

An optimal time between gain solutions can be defined by requiring balance between the errors in the $g_i(t)$ due to gain changes and the errors in the estimates of $g_i(t)$ due to finite signal-to-noise ratio. The condition for self-calibration to be possible is that *the time scale for gain changes should be much greater than the time taken for the noise per antenna to equal the source flux density.*

The errors in the estimated gains must feed back into the image and amplify the noise level. A noise analysis (Cornwell 1981) indicates that on a nearly unresolved source which is sufficiently strong for the errors in the gain estimates to be much less than a radian, the noise level in the background is increased by a small factor $\sqrt{\frac{N-1}{N-3}}$. The corresponding analysis cannot be performed for an extended source, but experience indicates that the noise level is seldom increased by more than a factor of 2 to 3.

5.4. Schwab's ℓ_1 and ℓ_2 solutions

Schwab (1981) has noted that minimization of sums of squares of errors (ℓ_2) is overly sensitive to spuriously discrepant points or outliers. He suggests that instead the ℓ_1 form should be minimized:

$$\mathcal{S} = \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) \left| \widetilde{V}_{ij}(t_k) - g_i(t_k) g_j^*(t_k) \widehat{V}_{ij}(t_k) \right| . \tag{10–16}$$

Tests on artificially generated data confirm the superiority of the ℓ_1 minimization algorithm when outliers are present. However, if the noise is normally distributed then the ℓ_2 minimization will, of course, provide superior results. Averaging of the data also alleviates this problem since seriously discrepant points are downweighted in the averages $\langle V_{ij}/\widehat{V}_{ij}\rangle$.

5.5. Spectral line self-calibration

In many spectral line observations the signal-to-noise in a single channel is too poor to allow separate self-calibration of each channel. Instead it is preferable to self-calibrate on the continuum emission and then use the gains so derived to correct the individual channel data. Note that separate bandpass calibration is required (see Lectures 5 and 17).

In cases where different lines appear at different locations, one could form a model having three dimensions, two of space and one of frequency, and then solve the corresponding least-squares problem,

$$\mathcal{S} = \sum_k \sum_l \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k, \nu_l) \left| \widetilde{V}_{ij}(t_k, \nu_l) - g_i(t_k) g_j^*(t_k) \widehat{V}_{ij}(t_k, \nu_l) \right|^2 . \qquad (10\text{–}17)$$

5.6. Spurious symmetrization

Suppose that we use a point source model for a slightly resolved source: if the number of array elements is sufficiently small, then the corrected phases will be significantly biased towards zero. As a consequence, after one iteration of self-calibration some features in the image will be seen reflected relative to the point-like component. However, if successive iterations are performed the spurious parts of the image will disappear.

Other, more subtle, symmetrizations are also possible but will disappear if enough iterations are performed. One example has been found by Linfield (1986): in simulations of the VLBA augmented by a high orbit satellite-based antenna, self-calibration failed to correct the gain of the orbiter. His explanation is that since one antenna is at one end of all the long spacings, it is difficult to distinguish between the astronomical structure phase, which is nearly equal on all spacings to the orbiter, and the antenna phase. Thus spurious symmetrization of the fine scale structure occurs. One cure is to calibrate the ground-based spacings internally before introducing the orbiter spacings, and then to allow only the orbiter phase to vary.

5.7. Non-convergence and non-uniqueness

Self-calibration nearly always converges to an answer but, especially for arrays (such as MERLIN) containing small numbers of elements, the final image is non-unique. As should now be apparent, there are a large number of free parameters available to the astronomer: apart from those inherent in the 'CLEAN' algorithm (see Lecture 8) the following can be altered in self-calibration:

1. Number of 'CLEAN' components passed in each iteration.
2. (u, v) range allowed for data to be used in solution.
3. Averaging time.
4. Type of solution and weighting scheme.

However, in most cases, poor choices for these and the 'CLEAN' parameters simply yield an image in which the effect of the corrections is not optimal. Only in cases of exceptionally poor (u, v) coverage (e.g., near declination 0°) and a relatively small number of array elements, ≤ 10, have two, or more, significantly different self-calibrated images been found in practice.

5.8. Baseline-related effects

If the gain errors are not purely element-based then self-calibration will, at some level, fail. The r.m.s. sidelobe level introduced by non-factorable errors is

$$\sigma_{B,C} = \frac{\sigma_{G,C}}{\sqrt{M}}, \tag{10–18}$$

where $\sigma_{G,C}$ is the r.m.s. baseline-related gain error and M is the number of such independent non-factorable errors. For the case of a reasonable synthesis with the VLA, $\sigma_{G,C} = 0.01$ and thus the best VLA image, in the absence of baseline-related calibration, will not have a dynamic range greater than about 35 dB.

Many different effects can lead to non-factorable gain errors. Clark (1981) has enumerated some of these and has described their correctability and relative magnitudes. We shall merely summarize some of these (see Clark's memorandum for further information):

1. Errors due to actual correlator problems. These are very unlikely in a digital correlator. They may be correctable if they are sufficiently constant with time.

2. Bandpass mismatches. These do not factor out on an antenna basis. They can, in principle, be corrected if the individual bandpasses are known. They are exacerbated by poorly adjusted delays.

3. Random, varying pointing errors. Simple self-calibration cannot correct for these if the size of the emission region is comparable to the main lobe of the primary beam $A(l, m)$ of the array elements.

4. Non-isoplanaticity of the atmosphere, i.e., different parts of the field of view to be imaged are seen through different cells in the atmosphere. Schwab (1984) has outlined a possible solution to this problem, which, owing to its complexity, has never been tried out; similar, but less unwieldy algorithms (Subrahmanya 1991) are under development for use in conjunction with the Giant Metrewave Radio Telescope (GMRT), now under construction in India.

5. Finite integration time and/or bandwidth. The latter can, in principle, be corrected, but this may be difficult to do in practice.

6. Incorrectly set sampling levels in the quantizers preceding the correlator.

7. Faulty analog quadrature networks.

All of these effects, save the first, are minimized by locating the source at the phase tracking center. The calibration and correction of baseline-based effects is discussed in Lecture 13.

6. Bibliography

A good and extensive review article on self-calibration appears in the 1984 Edition of the *Annual Review of Astronomy and Astrophysics* (Pearson & Readhead 1984).

References

Clark, B. G. 1981, VLA Scientific Memorandum No. 137, NRAO.

Cornwell, T. J. 1981, VLA Scientific Memorandum No. 135, NRAO.

Cornwell, T. J. & Wilkinson, P. N. 1981, *MNRAS*, 196, 1067–1086.

Cotton, W. D. 1979, *AJ*, 84, 1122–1128.

Dyson, F. J. 1975, *J. Opt. Soc. Am.*, 65, 551–558.

Ekers, R. D. 1984, in *Serendipitous Discoveries in Radio Astronomy*, Proceedings of a Workshop Held at the NRAO on May 4, 5, 6, 1983, K. I. Kellermann & B. Sheets, Eds., NRAO (Green Bank, WV), pp. 154–159.

Hamaker, J. P., O'Sullivan, J. D., & Noordam, J. E. 1977, *J. Opt. Soc. Am.*, 67, 1122–1123.

Jennison, R. C. 1953, *The Measurement of the Fine Structure of the Cosmic Radio Sources*, Ph. D. Thesis, University of Manchester.

Jennison, R. C. 1958, *MNRAS*, 118, 276–284.

Linfield, R. P. 1986, *AJ*, 92, 213–218.

Muller, R. A. & Buffington, A. 1974, *J. Opt. Soc. Am.*, 64, 1200–1210.

Pearson, T. J. & Readhead, A. C. S. 1984, *Ann. Rev. Astron. Astrophys.*, 22, 97–130.

Readhead, A. C. S., Walker, R. C., Pearson, T. J., & Cohen, M. H. 1980, *Nature*, 285, 137–140.

Readhead, A. C. S. & Wilkinson, P. N. 1978, *ApJ*, 223, 25–36.

Schwab, F. R. 1980, *Proc. S.P.I.E.*, 231, 18–25.

Schwab, F. R. 1981, VLA Scientific Memorandum No. 136, NRAO.

Schwab, F. R. 1984, *AJ*, 89, 1076–1081.

Subrahmanya, C. R. 1991 in *Radio interferometry: Theory, techniques, and applications*, Proceedings of the 131st IAU Colloquium, eds. T. J. Cornwell & R. A. Perley (San Francisco: ASP), 218–222.

Woolf, N. J. 1982, *Ann. Rev. Astron. Astrophys.*, 20, 367–398.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

11. Spectral-Line Observing I: Introduction

David J. Westpfahl
Physics Department, New Mexico Institute of Mining and Technology,
Socorro, NM 87801, U.S.A.

Abstract.
This lecture introduces the material necessary for successful planning, observing, and imaging of spectral-line data. This includes a review of important centimeter-wave spectral lines and the science done with them. It also includes choosing an array configuration, specifying velocity rest frames, and setting up the correlator. The lecture concludes with examples of an OH maser and of HI in emission and absorption.

1. Introduction

Two obvious ways to do spectroscopy with a radio telescope are illustrated in Figure 11–1. In multiband continuum mode a single frequency channel in each of several well-separated bands is observed to determine the overall shape of the spectrum of an object whose brightness changes slowly with frequency. The example in the figure shows observations of a star-forming region in the dwarf galaxy Holmberg II using the VLA (Tongue & Westpfahl 1995). The flux was determined at 335, 1465, and 4860 MHz (90, 20 and 6 centimeters or P, L, and C bands) for calculation of spectral indices.

In spectral line mode several adjacent frequency channels are observed within one band to determine the shape of the spectrum of an object whose brightness changes rapidly with frequency. The example in the figure shows observations of the 21 centimeter line of HI in M33 using the Synthesis Telescope of the Dominion Radio Astrophysical Observatory. M33 was detected in 173 channels, each of velocity width 1.64 km s^{-1}. Spectral line observing is the topic of this lecture.

Multiband-continuum mode is analogous to broadband photometry in the optical, spectral-line mode is analogous to spectroscopy. Spectral line observing is a distinct topic because of

1. the type of science done,
2. the way in which the correlator is used,
3. the additional practical considerations on planning observations, calibration, data display, and data analysis.

This lecture is organized around those three topics with examples.

1.1. The Goal of This Lecture

The goal of this lecture is to introduce the material necessary for successful planning, observing, and imaging of a spectral-line data cube with reasonable signal-to-noise ratio. This includes choosing an array configuration, choosing among the many correlator configurations, choosing and observing calibrators, allocating observing time between calibrators and science objects, and initial

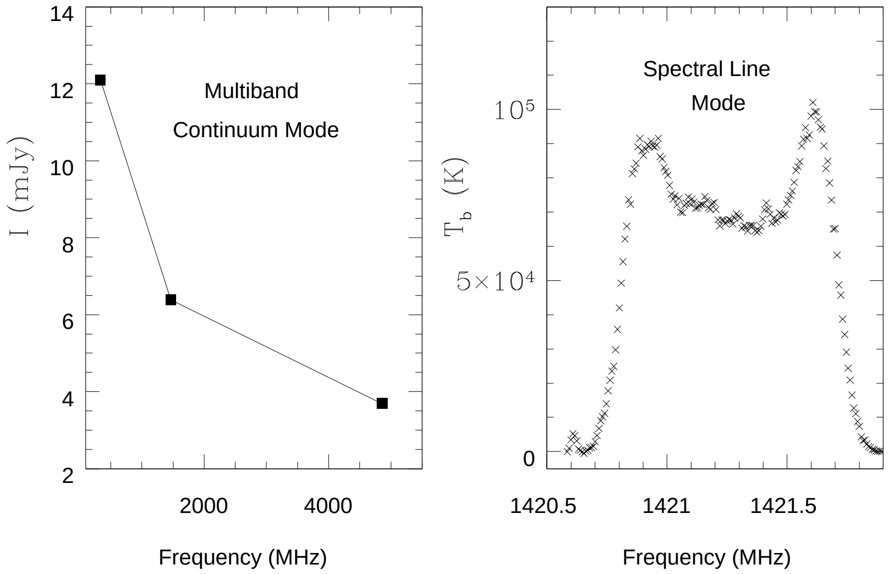


Figure 11–1. Examples of radio spectroscopy. On the left are multiband continuum observations of star-forming regions in Holmberg II made with the VLA. On the right is a spectral line observation of HI in M33 made with the Synthesis Telescope at Dominion Radio Astrophysical Observatory.

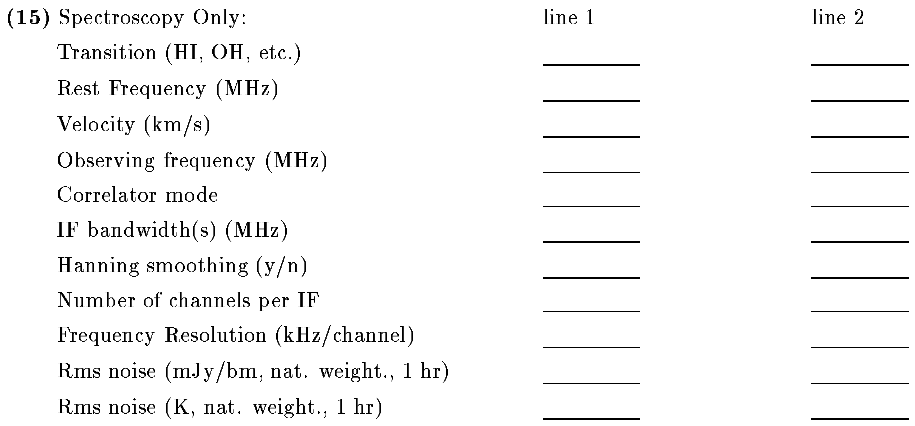
display of data cubes. Each observer must choose a scientific question, an object to study, a spectral line, and whether to observe it in emission or absorption.

While preparing this lecture I asked myself what is difficult about learning to do radio spectroscopy with an interferometer. The answer was, undoubtedly, understanding the information needed to fill in a proposal cover sheet and prepare an observe file for the first time. A survey of web sites and observer's manuals shows that all observatories with correlation spectrometers, on interferometers or single dishes, require similar information on their proposal cover sheets. Thus, this introduction to spectral-line observing will be built around the information you must gather to fill out an observing proposal cover sheet. This is the same information that you would need to prepare an observe file. Section 15 of the VLA proposal form, which is very similar to section 20 of the Australia Telescope Compact Array proposal form, is reproduced in Figure 11–2.

1.2. The Assumptions of this Lecture

I assume that you already know what a spectral line is. This saves me from having to define a spectral line myself.

I assume that the goal of observing and data reduction is to produce a cleaned data cube containing only the spectral line emission or absorption. That data cube will be the input to the data analysis step, in which the science product will be generated. Several stages of planning are required to assure that the correct data cube can be generated. Significant planning is required *before*



(15) Spectroscopy Only:	line 1	line 2
Transition (HI, OH, etc.)	________	________
Rest Frequency (MHz)	________	________
Velocity (km/s)	________	________
Observing frequency (MHz)	________	________
Correlator mode	________	________
IF bandwidth(s) (MHz)	________	________
Hanning smoothing (y/n)	________	________
Number of channels per IF	________	________
Frequency Resolution (kHz/channel)	________	________
Rms noise (mJy/bm, nat. weight., 1 hr)	________	________
Rms noise (K, nat. weight., 1 hr)	________	________

Figure 11–2. Section 15 from the VLA proposal cover sheet. The material in this lecture is an introduction to the decisions which must be made before filling in the blanks.

writing an observing proposal. In fact, several of the fields in observing proposal forms are there to be sure that you do the planning before writing the proposal. Thus, I assume that the Observational Status Summary and Guide for Spectral Line Observers (or similar materials if you are proposing to an observatory other than the NRAO) are companion documents to this lecture.

I assume that you wish to study the total intensity and not the polarization of the radio signal. This allows me to ignore the difficult topic of polarization, which is covered properly in Lecture 6.

Finally, I assume that the audience for this lecture is composed of newcomers to spectral-line projects. Thus, the references are chosen to be of maximum benefit to them and not to established experts. The purpose of this assumption is to prevent experts in the audience from arguing with me and each other over details.

2. Spectral-Line Science

After you have chosen your scientific problem and an object to observe you must choose a spectral line, then make sure that it falls within an available observing band (e.g., L, C, K, X, Q). The first two entries of Figure 11–2 ask for the line you plan to observe and its rest frequency. The material in this section provides overviews of the types of problems solved by spectroscopic observations and a few of the commonly-observed lines.

2.1. Why Use Spectral-Line Mode?

The obvious answer is to do spectroscopy, a better answer is to solve a particular scientific problem. Thanks to the success of the theory of atomic physics it is possible to calculate accurate frequencies and intrinsic widths for many radio lines and to calculate column density from an observed specific intensity. The Doppler shift allows observed frequency shifts to be interpreted as velocity. Thanks to the theory of radiative transfer it is possible to use observed specific intensity to determine optical depth and excitation (or spin) temperature. Thanks to kinetic theory we know that in regions dominated by collisions the excitation and kinetic temperatures are equal. These theories are summarized for the observer by Kerr (1968) and Verschuur (1974), among others.

The Doppler interpretation of frequency shift allows determination of the rotation and structure of the Milky Way and the systemic velocity and velocity fields of galaxies. These allow study of a wide range of topics from galactic structure to dark matter to the structure of the universe. Accurate line frequencies allow detection experiments to test theories of interstellar chemistry and primordial nucleosynthesis. Accurate column densities and temperatures allow study of the physical state of the interstellar medium, particularly the conditions necessary for star formation.

Spectral-line mode is also used to improve the quality of continuum observations. One obvious use would be to isolate radio frequency interference so that it could be eliminated from a continuum band. Another is to reduce the effects of bandwidth smearing. It is possible to measure Faraday rotation within one observing band in spectral-line mode if the object being studied has strong signal. Using spectral-line mode for continuum observing will be covered further in some of the other lectures in these proceedings.

2.2. Important Centimeter-Wave Spectral Lines

The first two entries in Figure 11–2 ask for the spectral line and rest frequency you plan to observe. This section briefly reviews commonly-observed lines and gives a few examples of the science done by observing them.

The 21-cm Line of Neutral Hydrogen and Other Hyperfine Lines Hydrogen has a spectral line at 21-cm because the proton and the electron can have their spins parallel or antiparallel. The small energy difference due to the magnetic interaction results in a long-lived state producing a radio line. This is known as hyperfine splitting of the ground state, and is also called a spin-flip transition. This spectral line is extremely important because hydrogen is common and because it occurs in the ground state, when the principal quantum number, n, is one. The successes of atomic physics allows determination of the number of hydrogen atoms in the ground state from the strength of the emission line (Kerr 1969, Verschuur 1974). The 21-cm line is easily detected in emission and absorption. The line can be used for Zeeman splitting experiments.

The 21-cm line was predicted by van der Hulst (1945) (available in English as reprinted by Sullivan 1982), first observed by Ewen & Purcell (1951), and quickly followed by Muller & Oort (1951) and Christiansen and Hindman, as reported by Pawsey (1951). Field (1959) developed the theory of the excitation of the HI line. This line has been of great importance for studying the rotation and ISM physics in the Milky Way and external galaxies. It is routinely used for the determination of rotation curves and searches for dark matter. Burton (1988) has reviewed the structure of the Milky Way determined from HI observations and Giovanelli & Haynes (1988) have reviewed the uses of observations of extragalactic HI.

Observed HI velocity dispersions are much larger than the intrinsic width of the line or of expected thermal line widths. This suggests that turbulence determines the observed width.

Deuterium has an analogous line at 92-cm and $^3\mathrm{He}^+$ has one at 3.46-cm. Both lines are important for testing the rate of nucleosynthesis in the early universe. There are no confirmed detections of the deuterium line. Deep integrations by Heiles, McCullough, & Glassgold (1993) led them to conclude that the interstellar medium is very efficient at forming the molecule HD, greatly depleting the abundance of atomic D. The hyperfine line of $^3\mathrm{He}^+$ has been studied by Balser et al (1994) and recently reviewed by Bania et al (1997).

Radio Recombination Lines As their name implies, radio recombination lines are found where ions can recombine with electrons, HII regions and planetary nebulae are examples. Recombination can result in the electron becoming bound in a state with large principal quantum number. Most of the time such a newly-bound electron jumps immediately to the ground state, resulting in emission in a resonance line. Occasionally the electron cascades downward from level to level, resulting in emission in a series of lines called recombination lines. Note that these are formed *after* recombination, not during recombination.

The lines are named using the element name, the lower principal quantum number, n, and its change, Δn. Consider hydrogen as an example. If the electron in a hydrogen atom jumps from level 111 to 110 the line is called H110α. If it jumps from 112 to 110 it is called H110β. Thus, $\Delta n = 1$ gives an α line, $\Delta n = 2$ gives a β line, $\Delta n = 3$ gives a γ line, etc. In this system the familiar optical Hα line would be called H2α.

The frequencies of the recombination lines of H and He^+ are especially easy to calculate theoretically and they are the recombination lines most often observed. For a hydrogenic atom the approximate line frequencies are

$$\nu = RcZ^2 \left[\frac{1}{n^2} - \frac{1}{(n + \Delta n)^2} \right] \tag{11–1}$$

where R is the Rydberg constant, c is the speed of light, and Z is the charge of the nucleus. The lines H110α, H138β, H158γ, H173δ, and H186ϵ all have frequencies near 4900 MHz in C band and are examples of regularly-observed recombination lines.

These lines are used to study the physical state, primarily pressure and electron temperature, of the gas in HII regions. They can be used in Zeeman splitting experiments to determine magnetic fields. Comparing the He and H recombination lines can give the abundance ratio of helium to hydrogen. HII regions may be optically thick to the Balmer lines, but optically thin to the radio lines, in which case the radio lines can be used to probe the internal structure of the HII region. HII regions often have strong continuum emission, so high dynamic range imaging may be required for good detection of the recombination lines. Recombination lines have been detected in planetary nebulae and nearby galaxies. Along with the lines of hydrogen and helium, radio recombination lines of carbon are well studied.

The first radio recombination line observed was H109α, at 5009 MHz, by Höglund & Mezger (1965). Their study of Galactic HII regions, Mezger &

Höglund (1967), remains a classic. Radio recombination line theory and observation have been reviewed by Gordon (1988).

Molecular Lines Molecules have lines in the centimeter radio bands because of their rotational and vibrational transitions. Shklovsky (1949) (retranslated by Sullivan 1982) predicted that CH, CH^+, and OH would have detectable radio lines. (In the same paper he calculated the first theoretical strengths of hyperfine lines.) The lines of OH, H_2O, NH_3, CS, SiO, methanol, and even redshifted CO are routinely observed with the VLA.

The first interstellar molecule discovered was OH, detected at 1667 MHz (18 cm) in absorption against Cassiopeia A by Weinreb, Barrett, Meeks, and Henry (1963). The late 1960's and early 1970's saw an explosion in molecular line detections, with hundreds of lines known today. Success in predicting line frequencies or measuring them in the laboratory has contributed to the rapid growth of the importance and number of known molecular lines.

Molecules are of overwhelming importance for studying dense, cold regions which do not have HI emission and are thick to optical lines. This is especially true for the envelopes of star-forming regions. By mass, about one half of the Milky Way's ISM is in molecular form, in some regions as much as 90% of the ISM is molecular.

Molecules allow study of astrochemistry. This is very different from terrestrial chemistry because of the low temperatures and pressures in the ISM. Much of interstellar chemistry appears to take place on the surfaces of interstellar grains. Turner (1974, 1988), and Turner & Ziurys (1988) have written reviews which are immediately useful for someone new to the field. Review volumes by Jorgensen (1993) and Nenner (1993) show how broad the field has become. Many large hydrocarbons are now known in the ISM, the field has gone beyond star formation to include a search for the building blocks of life. Miao et al (1994) are trying to detect amino acids.

Many molecules can be dissociated by optical light, so they are found only in cool regions where dust can screen them from background radiation. The lines of molecules in cold environments have little thermal broadening. This usually leads to the choice of narrow channels, particularly if it is necessary to resolve the shape of the line.

One of the most exciting fields of molecular line observations is the study of astronomical masers. OH, H_2O, SiO, HCN, and methanol are among the molecules known to form masers. The field is reviewed by Reid & Moran (1988).

3. Learn About the Object You Wish to Study

This step is so basic that experienced observers often forget that they do it. Before proceeding, learn the position and equinox, the size, the velocity, and at least an estimated brightness of the object you wish to observe. If you plan to observe objects of low velocity check to see if Milky Way emission is likely to be superimposed on them. Strong Milky Way signal may make an object undesirable as a science target. At the least the effect of Milky Way signal on the calibrators must be considered.

4. Choosing an Interferometer Configuration

Once you have chosen your science question, a radio source to observe, and a spectral line, you must choose an interferometer which best matches the size of the source you wish to observe. At the VLA this means choosing an array configuration. The object you wish to study may be too large for the VLA, in which case an interferometer with shorter baselines, such as the Super Synthesis Telescope at DRAO, would be a better choice. Similarly, your object may be too small, in which case a larger interferometer, such as Merlin or the VLBA, would be a better choice. Here are some guidelines.

Each interferometer or array configuration is sensitive to a limited range of angular scales determined by the baselines and the observing frequency or wavelength, λ. The largest observable angular scale in radians is $\frac{\lambda}{B_{\mathrm{short}}}$ and the smallest is $\frac{\lambda}{B_{\mathrm{long}}}$, where B_{short} and B_{long} are the shortest and longest baselines. The ratio $\frac{B_{\mathrm{long}}}{B_{\mathrm{short}}}$ is about 40 for each configuration of the VLA. The lengths of the baselines in each configuration are listed in the Observational Status Summary on the NRAO home page.

The value of $\frac{\lambda}{B_{\mathrm{short}}}$ should equal or exceed the angular size of *the largest structure in any single channel.* If the source that you are observing has velocity gradients (e.g., a rotating galaxy) then $\frac{\lambda}{B_{\mathrm{short}}}$ may be smaller than the diameter of the source provided that is is as large as the largest extent of the source in any one channel. This can make choosing a configuration for a spectral-line project different from that for a continuum project.

This standard requires you to have a good idea of the size and velocity structure of an object before you observe it, which causes a chicken-and-egg problem if those are the properties you wish to determine. The sizes of many objects can be estimated from existing single-dish observations or from their optical sizes.

Many objects have structure on a large range of sizes, so large that a single array configuration or a single interferometer might not be sensitive to them all. Such objects may be observed with more than one array configuration at the VLA or with more than one interferometer, e.g., the VLA and Merlin.

As an array gets larger its synthesized beam gets smaller, surface brightness sensitivity is reduced, and more observing time is needed to detect a given column density of gas. Thus, it is usually easiest to detect broadly-spread emission (such as the HI in a nearby galaxy) with the smaller array configurations which have good surface brightness sensitivity.

Absorption is seen against a background source which is usually small. In most absorption projects detecting the emission is undesirable. Thus, absorption projects are well served by the larger array configurations which resolve out the emission and whose synthesized beams are well matched to the size of the background source.

5. Choosing an Observing Frequency

The third and fourth entries of Figure 11–2 require that you relate the observed frequency of a spectral line to the velocity of the object being observed. This

Table 11–1. Velocity Rest Frames Derived from the Topocentric System

Rest Frame	Corrected for	Amplitude of Correction, km s^{-1}
Geocentric	Earth rotation	0.5
Earth-Moon Barycentric	Effect of the Moon on the Earth	0.013
Heliocentric	Earth's orbital motion	30
Solar System Barycentric	Effect of planets on the Sun	0.012
Local Standard of Rest	Solar motion	20
Galactocentric	Milky Way rotation	230
Local Group Barycentric	Milky Way motion	~100
Virgocentric	Local Group motion	~300
Microwave Background	Local Superclucster motion	~600

turns out to be more difficult than you might expect because we live on a moving Earth in a moving solar system.

5.1. Velocity Reference Frames

Most radio telescopes are attached to the surface of the Earth, so they are in the topocentric rest frame. Most astronomical objects are not. To place the desired spectral line in the central channel of the total bandwidth it is essential that you specify the velocity and rest frame of the object you wish to observe. The control systems of most radio telescopes can transform among several rest frames by allowing for the Earth's rotation and orbital motion, the solar motion, and center of mass effects within the solar system. The observing frequency is then tracked to keep a spectral line centered in a specific channel. Without this tracking the spectral line would drift from channel to channel, smearing the line. The tracked frequency is usually called the sky frequency.

Velocity reference frames start with the topocentric system and, by successive transformations, work toward the Virgocentric system. Their relationships are summarized in Table 11–1. Correcting the topocentric system for the Earth's rotation transforms to the Geocentric system. This correction is no larger than 0.5 km s^{-1}, is zero in the direction of the celestial poles, and reaches a maximum and minimum on the celestial equator. Correcting for the very small motion of the Earth relative to the Earth-Moon barycenter transforms to the Earth-Moon barycentric system, an additional correction for the orbital motion of the Earth-Moon barycenter transforms to the heliocentric system. Correcting for the motion of the barycenter of the solar system transforms to the Solar System barycentric system. This system is often incorrectly called the heliocentric system, the difference between the two is small so the confusion is forgivable. The barycentric system is commonly used in observing galaxies, the barycentric velocity is listed in many catalogs. Velocities of objects in the Milky Way are often cataloged relative to the local standard of rest or LSR. Correcting the barycentric velocity for the solar motion transforms to the LSR. Correcting for Milky Way rotation transforms to the galactocentric system. The data needed for these transformations have been standardized by the International Astronomical Union and are available in the IAU Reports on Astronomy (Appenzeller 1997) and included in the Astronomical Almanac (Seidelman 1992).

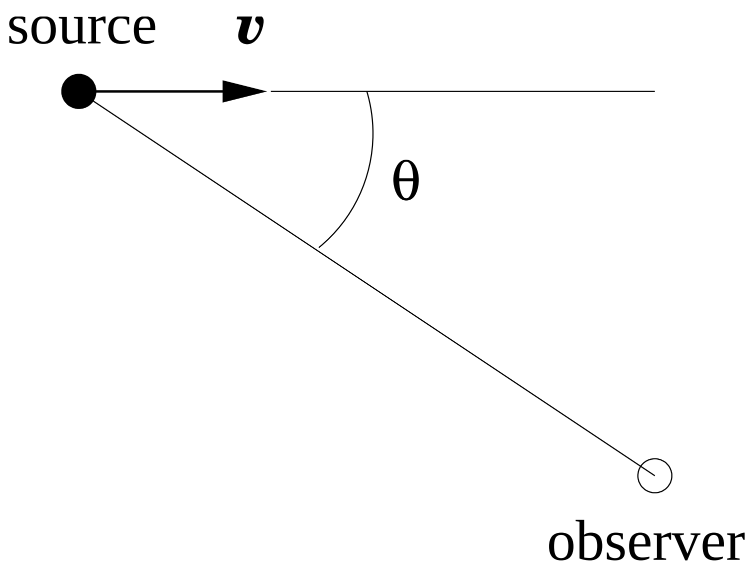


Figure 11–3. The geometry for the Doppler shift calculation. A source moves with velocity **v** which makes an angle θ relative to the line of sight of the observer.

There is a transformation to the barycenter of the Local Group, and, subsequently, a correction for the motion of the Local Group toward the Virgo cluster, leading to the Virgocentric system. There is even a transformation to the rest frame of the cosmic microwave background radiation. These reference frames are not accommodated by the telescope control system, but they do occasionally appear in catalogs. Several reviews of these transformations can be found in the conference volume edited by Madore & Tully (1986).

5.2. The Relationship Between Frequency and Velocity

Once the velocity and velocity rest frame of the object and the rest frequency of the desired spectral line have been specified the on-line system can relate the sky frequency to the topocentric frequency. Consider an object moving at velocity **v** relative to an observer, as in Figure 11–3. Let the object emit a spectral line of rest frequency ν_0. The velocity, rest frequency, and received frequency, ν, are related by

$$\nu = \nu_0 \frac{\sqrt{1 - \frac{v^2}{c^2}}}{1 - \frac{v}{c}\cos\theta} \tag{11–2}$$

where θ is the angle between the direction of **v** and the line of sight. If the object is moving radially away from the observer then $\theta = \pi$ and

$$\nu = \nu_0 \frac{\sqrt{1 - \frac{v}{c}}}{\sqrt{1 + \frac{v}{c}}} \,. \tag{11–3}$$

Inverting,

$$\frac{v}{c} = \frac{\nu_0^2 - \nu^2}{\nu_0^2 + \nu^2} \,. \tag{11–4}$$

Equations 11–2, 11–3, and 11–4 are relativistically correct statements of the Doppler shift.

There are two commonly-used methods for approximating the result in Eq. 11–4 starting with Eq. 11–3 when v is much smaller than c. In the radio velocity definition the approximation is made directly with frequencies, giving

$$\frac{v_{\rm radio}}{c} = \frac{\nu_0 - \nu}{\nu_0}. \tag{11–5}$$

In the optical velocity definition the frequencies are replaced with wavelengths and the approximation is made, giving

$$\frac{v_{\rm optical}}{c} = \frac{\lambda - \lambda_0}{\lambda_0} \equiv z\,, \tag{11–6}$$

where z is the redshift. In terms of frequency this becomes

$$\frac{v_{\rm optical}}{c} = \frac{\nu_0 - \nu}{\nu}. \tag{11–7}$$

Both velocity systems are in use, with the optical definition being more common. Read catalog headings carefully! It is worth noting that

$$\frac{v_{\rm radio}}{v_{\rm optical}} = \frac{\nu}{\nu_0} \tag{11–8}$$

so Eq. 11–8 can be used to convert from one system to the other.

6. Choosing a Configuration of the Spectroscopic Correlator

The fifth through ninth entries of Figure 11–2 require that you specify how you wish to use the spectroscopic correlator. Knowing a little about how the correlator works allows you to understand what choices you are making and why.

Conceptually, the simplest way to do spectroscopy is with filter banks, as in Figure 11–4. In the case of a simple, two-element interferometer the telescopes would have identical filter banks. The output of each filter in the bank of one telescope would be correlated with the output of the corresponding filter in the bank of the other telescope. This would give the visibility as a function of u, v, and frequency, $V(u, v, \nu)$. Many radio telescopes have worked this way. The drawback to this method is that many filter banks are required if the telescope is to accommodate a variety of science goals in many frequency bands.

6.1. Using Lags to Define a Fourier Transform Relationship Between Time and Frequency

Instead of the filter bank the correlator can provide several frequency channels within an observing band by using the same Fourier techniques which allow imaging. This is one of the important differences between spectral line and continuum modes. This section presents the big picture only, the details are properly worked out in Lecture 4.

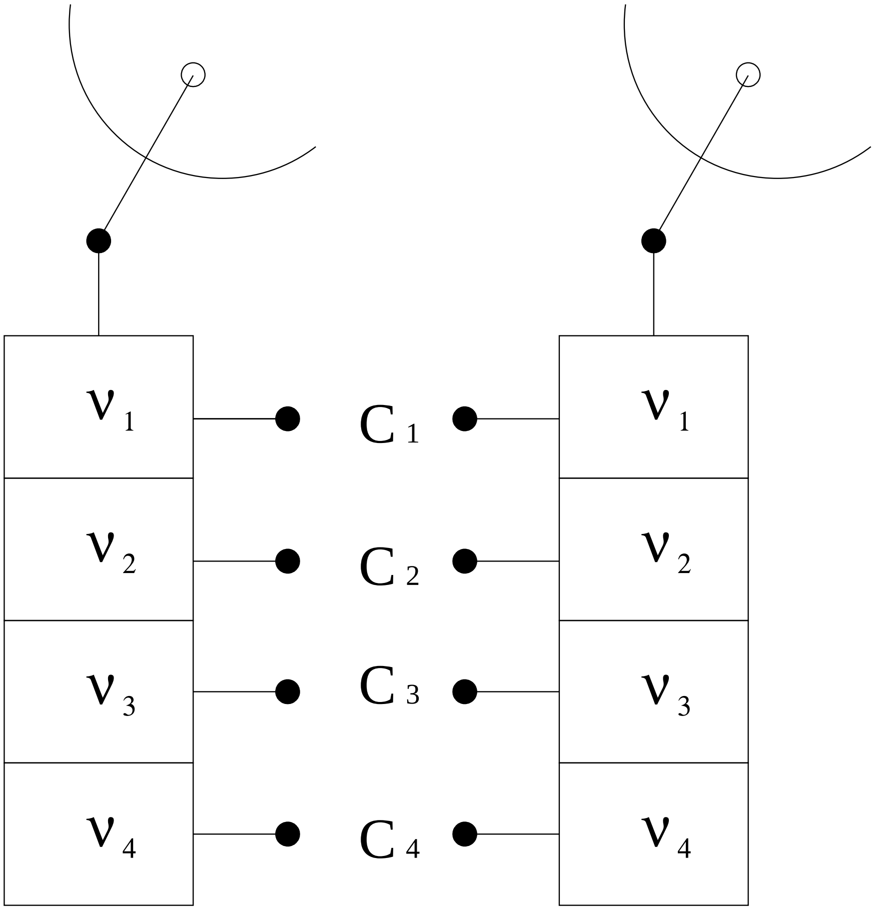


Figure 11–4. Spectroscopy using filter banks. Each telescope has a four-channel filter bank, the outputs of corresponding filters are sent to a correlator. This type of spectroscopic telescope must have many filter banks if it is to support a variety of science projects.

We wish to have the visibility as a function of frequency, $V(u, v, \nu)$, for several frequency channels of known center frequency and separation. The number of channels and their centers should be adjustable to accommodate a wide range of science objectives.

The correlator is already good at inserting and recording time delays in the data paths to compensate for the geometric delay and to stop the fringes. Thus, it is straightforward to have the correlator record the visibility as a function of time. Furthermore, the visibility can be shifted in time by copying the signal (or using a shift register). The shift in time is called the lag, τ. Thus, the correlator can record the visibility as a function of lag, $V(u, v, \tau)$.

A visibility can be copied many times with many different lags. Ideally an infinite number of lags, positive and negative, could be applied, making $V(u, v, \tau)$ a continuous function of τ. It would then be possible to define a

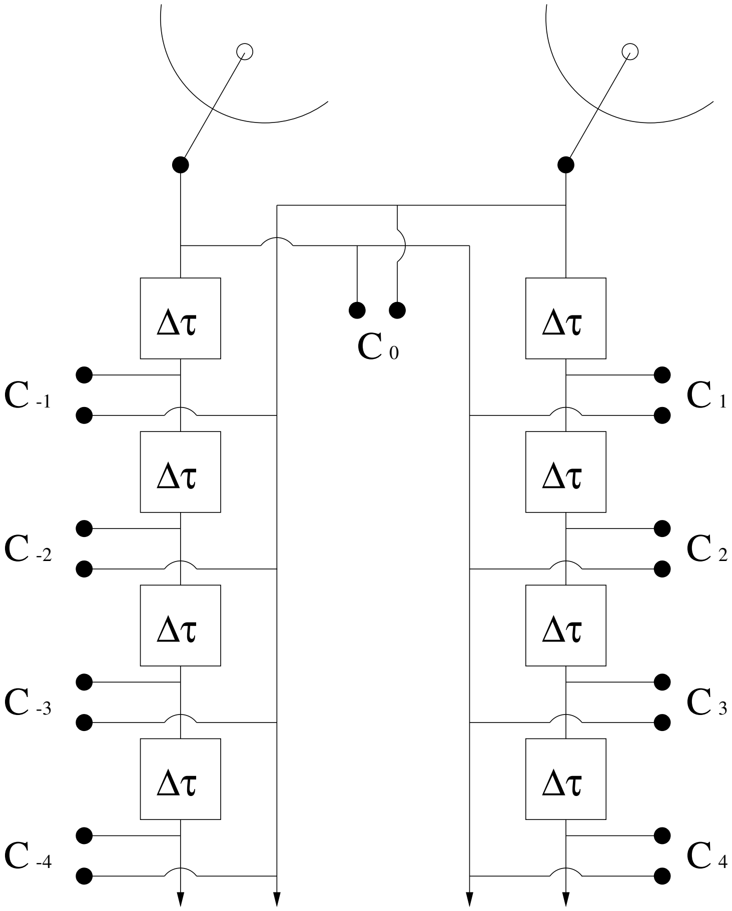


Figure 11–5. Conceptual layout of a spectroscopic correlator which uses a fixed lag interval, $\Delta\tau$. The individual correlators, C, give the visibility as a function of τ. The subscripts on the C's give the number of lag intervals applied to the left telescope relative to the right telescope.

Fourier transform relationship between τ and ν,

$$V(u,v,\nu) \;=\; \int_{-\infty}^{+\infty} V(u,v,\tau)e^{-2\pi i\nu\tau}\;d\tau\,. \tag{11–9}$$

If the telescope and receiver could be sensitive to all frequencies it would be possible, in principle, to determine the visibility, hence the sky brightness, at all frequencies! The drawback is that this would require an infinite amount of time.

6.2. Sampling

We do not need to know $V(u,v,\nu)$ for all values of ν, only for those in a limited part of the radio spectrum. Similarly, we do not need channels of infinitesimal separation, small but finite separation will do nicely. Thus, the information that we want can be found using a finite number of lags and a discrete Fourier transform. It is especially convenient to sample the visibility at some fixed lag interval, $\Delta\tau$. One-half of its inverse, $\frac{1}{2\Delta\tau}$, defines the total frequency bandwidth of the observation.

We wish to know $V(u, v, \nu)$ for each of N channels. The visibility is a complex quantity, so we must specify two quantities for each visibility, or a total of $2N$ quantities. We must then measure at least $2N$ quantities to determine the N visibilities.

We might choose to measure all lags between $-N\Delta\tau$ and $+(N-1)\Delta\tau$, a total of $2N$ lags (a lag of zero is included). One-half of the inverse of the maximum lag, $\frac{1}{2N\Delta\tau}$, defines the separation of the frequency channels, $\Delta\nu$. Thus, the sampling interval and the total number of samples defines the number and separation of the channels.

Such a correlator is shown conceptually in Figure 11–5. The signals from the telescopes are correlated with zero delay, C_0, and with several positive and negative delays, C_1, C_{-1}, etc., in which the subscript gives the number if delay intervals, $\Delta\tau$.

Most spectroscopic correlators produce channels of equal frequency separation. This means that the channels will not be of equal velocity separation. If the channels are small and if there are not too many this inequality can usually be ignored, but if the overlapping channels in two data sets with different center frequencies are to be found it is easier to find them in frequency than velocity.

6.3. The Gibbs Phenomenon and Hanning Smoothing

The finite width of the lag function has the advantage of leading to a finite total bandwidth but the disadvantage that we synthesize a square band. This leads to ringing at the beginning and end of the band, and in response to any sharp spectral features, similar to the Gibbs phenomenon familiar from Fourier series. The regions of ringing at the ends might be avoided by doubling the number of lags to double the number of channels and then recording only the central fifty percent of the channels in the band, discarding the edges. That seems wasteful. A more attractive solution may be to smooth the data in frequency space. Several smoothing functions exist, the one used most often is Hanning smoothing. Consider an example in which we want N channels with Hanning smoothing. In the correlator the number of lags actually used is double the number needed, the channel width is halved but the number of channels is doubled to $2N$ to allow full coverage of the observing band. Visibilities are obtained for these $2N$ channels. The visibilities are then smoothed and recorded in N channels. The visibilities in the first three of the $2N$ channels are averaged with weights of $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ and the result stored in the first of the N channels. The third, fourth, and fifth of the $2N$ channels are similarly averaged and the result recorded in the second of the N channels, the fifth, sixth, and seventh are averaged and recorded in the third, and so on. This very effectively reduces the ringing, allowing all N channels to be kept. The effect of Hanning smoothing on a Fourier series approximation to a square wave is shown in Figure 11–6.

6.4. Channel Separation versus Channel Bandwidth

The center-to-center separation of the channels and the width of the channels are not necessarily the same, and smoothing can change their relationship. Sampling over a limited time range is equivalent to multiplying all possible samples with a

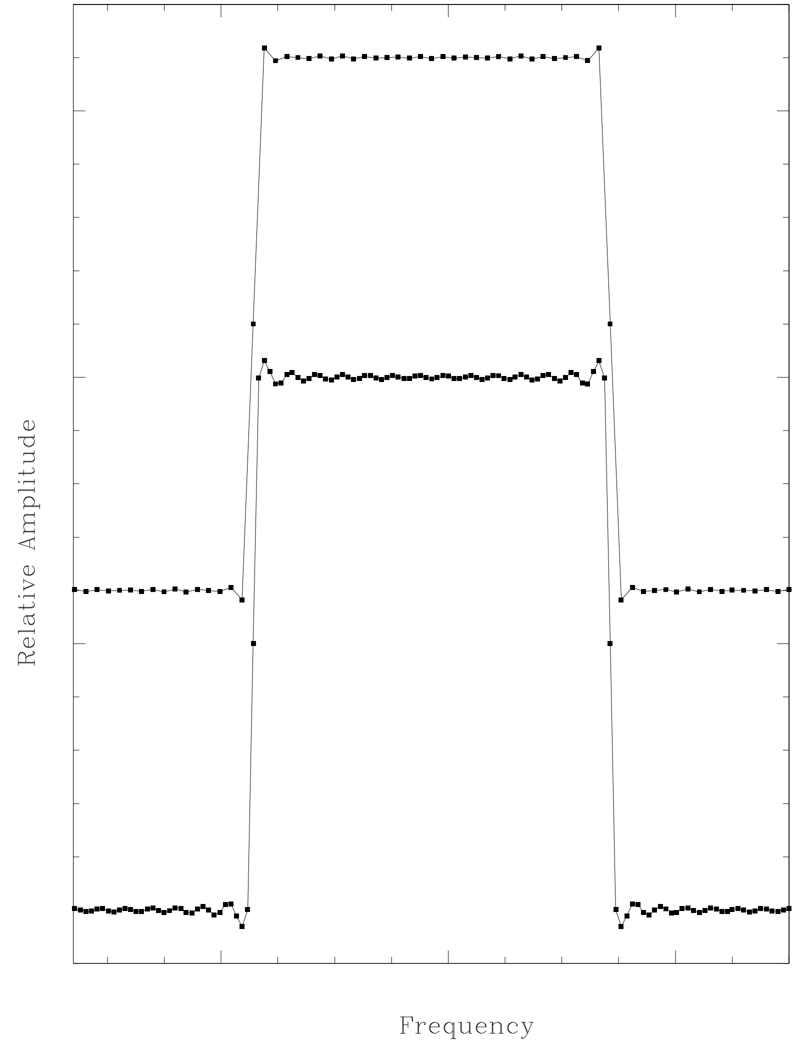


Figure 11–6. A square wave of unit amplitude approximated by 100 terms of a Fourier series, bottom, and the approximation after Hanning smoothing, top. Each point represents the center of a channel. The smoothed curve has been arbitrarily shifted vertically so it does not overlap with the original curve. Hanning smoothing greatly reduces the overshoot at the discontinuities and almost entirely removes the ripple in the middle of the frequency band.

rectangular function of finite width. The Fourier transform of this product is the convolution of the true spectrum with the Fourier transform of the rectangular function, which is a sinc function. Thus, each channel has the width of the sinc function. The details have been worked out by Thompson, Moran, & Swenson (1991). In the absence of Hanning smoothing a typical correlator produces

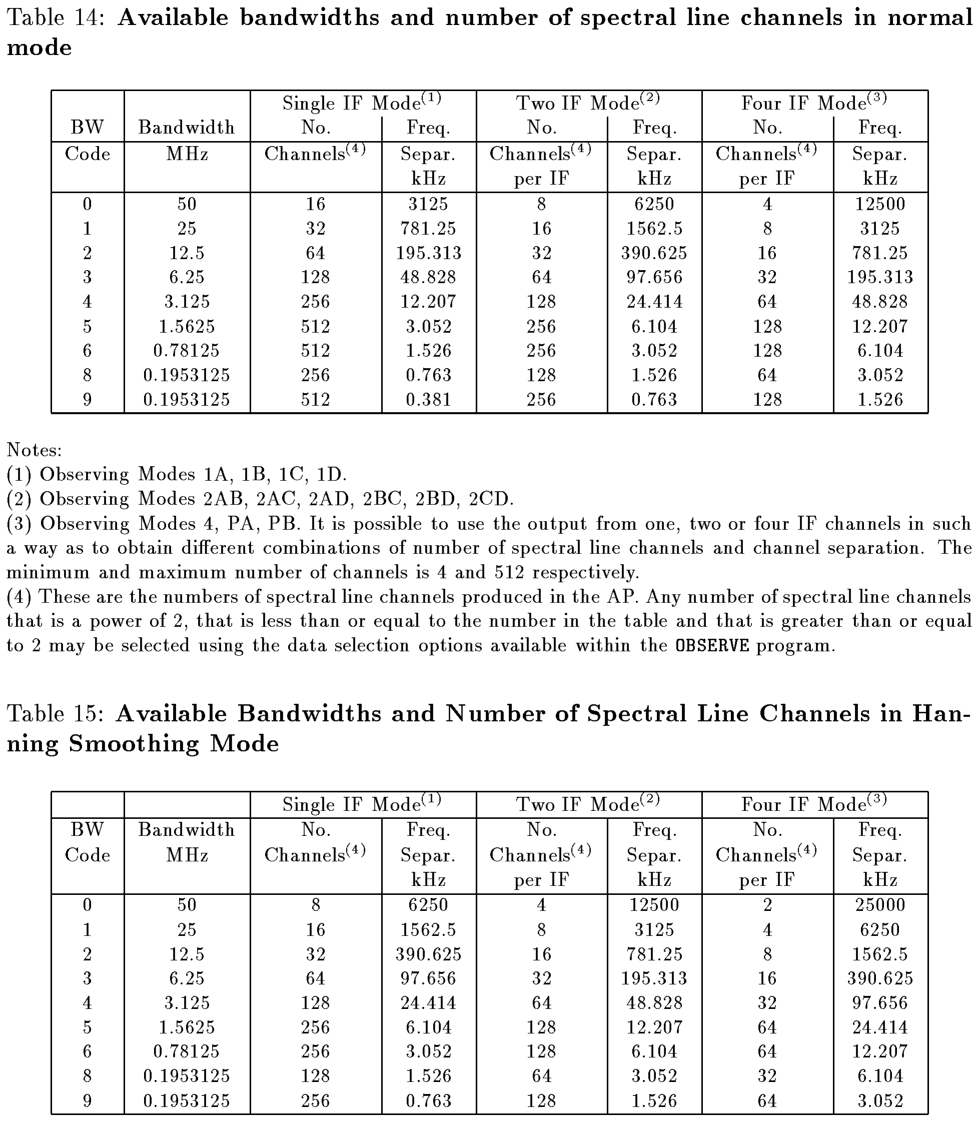


Table 14: **Available bandwidths and number of spectral line channels in normal mode**

		Single IF Mode$^{(1)}$		Two IF Mode$^{(2)}$		Four IF Mode$^{(3)}$	
BW	Bandwidth	No.	Freq.	No.	Freq.	No.	Freq.
Code	MHz	Channels$^{(4)}$	Separ.	Channels$^{(4)}$	Separ.	Channels$^{(4)}$	Separ.
			kHz	per IF	kHz	per IF	kHz
0	50	16	3125	8	6250	4	12500
1	25	32	781.25	16	1562.5	8	3125
2	12.5	64	195.313	32	390.625	16	781.25
3	6.25	128	48.828	64	97.656	32	195.313
4	3.125	256	12.207	128	24.414	64	48.828
5	1.5625	512	3.052	256	6.104	128	12.207
6	0.78125	512	1.526	256	3.052	128	6.104
8	0.1953125	256	0.763	128	1.526	64	3.052
9	0.1953125	512	0.381	256	0.763	128	1.526

Notes:
(1) Observing Modes 1A, 1B, 1C, 1D.
(2) Observing Modes 2AB, 2AC, 2AD, 2BC, 2BD, 2CD.
(3) Observing Modes 4, PA, PB. It is possible to use the output from one, two or four IF channels in such a way as to obtain different combinations of number of spectral line channels and channel separation. The minimum and maximum number of channels is 4 and 512 respectively.
(4) These are the numbers of spectral line channels produced in the AP. Any number of spectral line channels that is a power of 2, that is less than or equal to the number in the table and that is greater than or equal to 2 may be selected using the data selection options available within the OBSERVE program.

Table 15: **Available Bandwidths and Number of Spectral Line Channels in Hanning Smoothing Mode**

		Single IF Mode$^{(1)}$		Two IF Mode$^{(2)}$		Four IF Mode$^{(3)}$	
BW	Bandwidth	No.	Freq.	No.	Freq.	No.	Freq.
Code	MHz	Channels$^{(4)}$	Separ.	Channels$^{(4)}$	Separ.	Channels$^{(4)}$	Separ.
			kHz	per IF	kHz	per IF	kHz
0	50	8	6250	4	12500	2	25000
1	25	16	1562.5	8	3125	4	6250
2	12.5	32	390.625	16	781.25	8	1562.5
3	6.25	64	97.656	32	195.313	16	390.625
4	3.125	128	24.414	64	48.828	32	97.656
5	1.5625	256	6.104	128	12.207	64	24.414
6	0.78125	256	3.052	128	6.104	64	12.207
8	0.1953125	128	1.526	64	3.052	32	6.104
9	0.1953125	256	0.763	128	1.526	64	3.052

Figure 11–7. Tables of spectral line modes of the VLA correlator as they appear on the NRAO home page. Before writing a proposal one must choose normal mode or Hanning smoothing, the bandwidth code, which specifies the channel bandwidth and the total bandwidth, and the number of IFs.

channels whose full width at half maximum is 1.2 times their separation. With Hanning smoothing the channel width is equal to the channel separation.

6.5. An Example, the VLA Correlator

The correlator configurations available at the VLA are shown in Figure 11–7, which is taken directly from the NRAO home page. Notice that the bandwidth and channel separation change by factors of two. This is because the correlator uses discrete Fourier transforms. The observer must choose between normal mode and Hanning smoothing, choose the channel bandwidth, the total bandwidth, the number of IFs, and which IFs to use.

Two of these choices are relatively easy. Hanning smoothing is preferred over normal mode because a larger fraction of the total bandwidth is usable when the Gibbs phenomenon is avoided. Choose Hanning smoothing unless there is a specific reason for not smoothing, e.g., bandwidth considerations. Alternatively, Hanning smoothing can be done after observing, as part of data reduction.

Similarly, two IFs are preferred over one IF. Two IFs give more information than one because they record different polarizations. More information means more sensitivity in a given amount of observing time. Two IFs are preferred unless there is a specific reason for choosing only one, which could be because of frequency resolution. Four IFs do not give more information than two IFs unless the additional IFs have a different frequency center or bandwidth.

The choice of the correlator mode is also straightforward. At the VLA the correlator has two stages of mixing, so there are four IFs, labeled A, B, C, and D. IFs A and B contain the right circularly polarized signal, C and D the left. To specify the correlator mode you must choose the number of IFs (1, 2, or 4) and the specific IFs you wish to use. Choose the IFs on the basis of polarization. If you are using one IF decide ahead of time which polarization you want. If you are using two IFs for more sensitivity choose a mode which includes both polarizations, 2AC is the standard, but 2AD, 2BC, or 2BD would also include both polarizations. If you are using four IFs then you automatically use all the IFs, so there is nothing to choose, the mode is simply 4. Lecture 6 describes how to use the modes for detecting polarization.

In many spectral line projects the channel bandwidth is chosen by examining the desired velocity resolution and sensitivity. A detection experiment might use relatively wide channels matched to the total expected width of the spectral line, perhaps 20 or 40 km s^{-1}, for good sensitivity, while observations of a maser would use narrow channels, perhaps less than 1 km s^{-1}, because maser lines are intrinsically narrow. As a guideline, consider whether you need to resolve the shape of the spectral line or merely detect its presence, then choose the channel bandwidth accordingly. Choosing the channel bandwidth is equivalent to choosing the total number of lags that the correlator will use.

In most projects the total bandwidth is chosen to allow several line-free channels on either side of the spectral line. These channels are used to detect the continuum radiation so it can be subtracted, leaving only the line signal. This choice, too, is often based on velocity. If you wish to study the rotation of a galaxy in HI the total width of the line could be more than 300 km s^{-1}. The total bandwidth must be significantly wider than this to allow for detection and subtraction of the continuum. It is important to remember that even with Hanning smoothing some channels at each end of the band will be unusable. This could eliminate as many as twenty percent of the channels, ten percent on each end. Allow a generous total bandwidth.

Choosing the total bandwidth is equivalent to choosing the sampling interval. Choosing both the channel bandwidth and the total bandwidth fixes the number of channels.

Occasionally the choice of channel bandwidth and total bandwidth are in conflict so that there is not a correlator mode which can meet both requirements. Let the science determine which requirement can be relaxed.

7. Sensitivity and Calibration

The last two entries in Figure 11–2 require that you understand the sensitivity, in specific intensity and brightness temperature, of your proposed observations. These can be used as a guide to determine the total amount of time you must observe the radio source you wish to study and how much time must be dedicated to calibration. In this section I_ν is specific intensity and ΔI is sensitivity in the same units.

7.1. Sensitivity

The sensitivity formula for the rms noise in milli-Janskys per beam in a single spectral channel observed with the VLA and imaged with natural weighting is

$$\Delta I_{\mathrm{rms}} = \frac{K}{\sqrt{N(N-1)\Delta t_{\mathrm{hr}}\Delta\nu_{\mathrm{MHz}}}} \text{ mJy beam}^{-1} . \qquad (11\text{–}10)$$

N is the number of antennas, n is the number of IFs, Δt_{hr} is the time on source in hours, $\Delta\nu_{\mathrm{MHz}}$ is the channel bandwidth in MHz, and K is a constant which depends upon the observing band (given in the Observational Status Summary). This can be compared with an estimated brightness to determine the amount of time needed to detect a given source at a given signal-to-noise ratio. In many cases the estimated brightness is not known in mJy per beam, but rather in brightness temperature or column density.

Conversion to Brightness Temperature Conversion to brightness temperature is done using Planck's blackbody law,

$$I_\nu(T) = \frac{2h\nu^3}{c^2}\frac{1}{e^{\frac{h\nu}{kT}}-1} \qquad (11\text{–}11)$$

where I_ν is the specific intensity, T is temperature, ν is frequency, and h, k, and c are familiar constants. At low frequencies $h\nu \ll kT$. Using a Taylor expansion for the exponential gives

$$e^{\frac{h\nu}{kT}} - 1 = 1 + \frac{h\nu}{kT} - 1 = \frac{h\nu}{kT} . \qquad (11\text{–}12)$$

The black body law becomes

$$I_\nu = \frac{2h\nu^3}{c^2}\frac{kT}{h\nu} = \frac{2\nu^2 kT}{c^2} = \frac{2kT}{\lambda^2} , \qquad (11\text{–}13)$$

which is the Rayleigh-Jeans Law. Inverting gives

$$T_b = \frac{c^2}{2k\nu^2} I_\nu = \frac{\lambda^2}{2k} I_\nu \qquad (11\text{–}14)$$

which is the definition of brightness temperature. Note that I_ν is usually observed in units of Janskys beam^{-1}, so it must be converted to Watts m^{-2} Hz^{-1} steradian^{-1} before calculating the brightness temperature. For data from the VLA reduced with AIPS the AIPS header will contain the full width at half maximum of the clean beam, which can be used to make the conversion.

Conversion to Column Density The conversion to column density is different for each spectral line because the emission mechanisms, transition probabilities, statistical weights, and oscillator strengths are different. The conversion for the 21-cm line of hydrogen is worked out by Kerr (1969) and Verschuur (1974). If the spin temperature is constant along the line of sight then the number of atoms in the ground state per unit frequency interval (1 Hz) in a column of cross section 1 cm^2, n_ν, is

$$n_\nu = 3.88 \times 10^{14} T_s \tau_\nu \tag{11–15}$$

where T_s is the spin or excitation temperature and τ_ν is the optical depth as a function of frequency. In velocity units the number of atoms in a unit interval of velocity (1 km s^{-1}) is

$$n_v = 1.82 \times 10^{18} T_s \tau_v \,. \tag{11–16}$$

Multiplication of n by the width of the line, in Hertz or km s^{-1} as appropriate, completes the conversion to column density.

7.2. Calibration

To use observing time effectively you must have a plan for adequate, but not excessive, observation of calibrators. This allows enough observing time on the science object for good signal-to-noise ratio given the limitations of finite observing time.

Spectral-line observations must include at least one primary flux calibrator and at least one phase calibrator located close to the object being studied, just as in continuum observing. Spectral-line projects require an additional calibration step to make sure that the response is the same in all channels. This is called bandpass calibration. It is done by observing a source whose spectrum is known, ideally it would be flat. The known shape of the spectrum is then used to normalize the sensitivities of the channels. Some good bandpass calibrators for L band at the VLA are 3C48, 3C147, and 3C286, which are primary flux calibrators as well.

At the risk of stating the obvious, the correlator setup and central velocity should be exactly the same for the object and all the calibrators. There are occasional exceptions to this, but they are few.

The bandpass calibrator should be observed long enough so that bandpass calibration does not make a significant contribution to the signal-to-noise ratio in the final data cube. The sensitivity formula can be applied using the width of a single channel. Choose an observing time large enough so that the signal-to-noise ratio of the calibrator significantly exceeds the expected signal-to-noise ratio of the object to be studied. Notice that this requires knowledge of the brightness of the object, just the quantity that you usually want to determine! It is usually possible to estimate the brightness from catalogs or published observations. In some cases it may be necessary to obtain a single-dish observation well ahead of time for planning purposes.

Consider an example. According to the sensitivity formula the noise is proportional to a constant times $(\sqrt{N(N-1)n\Delta t_{\mathrm{hr}}\Delta\nu_{\mathrm{MHz}}})^{-1}$. N, n, $\Delta\nu$, and the constant will be the same for the object and the bandpass calibrator, so our requirement on the signal-to-noise ratio is that $I_{\mathrm{cal}}\sqrt{t_{\mathrm{cal}}}$ must be significantly greater than $I_{\mathrm{obj}}\sqrt{t_{\mathrm{obj}}}$. Let the bandpass calibrator have a flux density of 7.0 Janskys and let the object have an estimated brightness of 100 milliJanskys. If the object is observed for four hours then equal signal-to-noise ratios would be obtained if the bandpass calibrator is observed for $\frac{0.4}{7.0}$ hours or about 3 minutes. To have the signal-to-noise ratio of the bandpass calibrator be 10 times that of the object would require about 300 minutes, which is good from a calibration standpoint, but probably would take too large a fraction of the project time. A signal-to-noise ratio four times that of the object would require about 48 minutes, which is a good compromise between time and good calibration.

8. Data Reduction and Analysis

I assume that your observations have gone well, with no instrumental problems and a minimum of radio frequency interference. The data must undergo calibration, imaging, cleaning, and continuum subtraction. These topics are well covered in the AIPS Cookbook and other NRAO materials. Here I introduce data cubes and continuum subtraction.

8.1. Data Cubes

Each frequency channel in the spectral-line data set is imaged using the same procedures as would be applied to the single channel in a continuum data set. The channels can be used individually, but usually are assembled into a single data structure with three axes – right ascension, declination, and frequency – called a data cube, Figure 11–8 contains an example. Mathematically a data cube is is not a true cube, its axes have two different types of units, angle and frequency, and it does not necessarily have the same number of pixels on each axis. None the less, the description cube is conceptually convenient.

Cubes are convenient for storing, displaying, and manipulating the spectral data. Examples are given in the next section.

8.2. Continuum Subtraction

Our goal is a cleaned data cube from which the continuum has been subtracted. To subtract the continuum it is first necessary to know which channels contain only continuum and which contain both the continuum and the line. The channels containing only continuum are called line-free channels. Distinguishing the line-free channels from those containing line signal is usually done by making an initial data cube expressly for this purpose. If the line signal is very weak compared to the continuum it is usually necessary to clean this cube, if the line signal is strong cleaning may be unnecessary.

The continuum signal can be subtracted in the map plane or in the uv data. If in the map plane then the order of reduction steps is imaging, averaging the

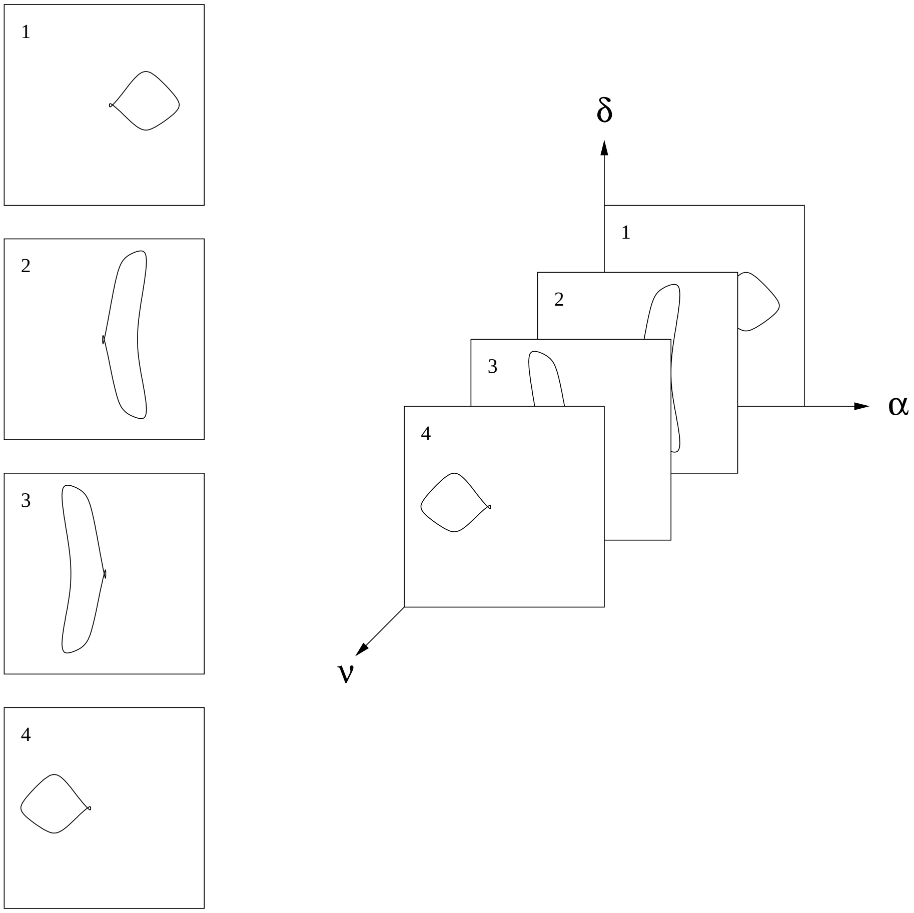


Figure 11–8. An example of four spectral line channels, on the left, which are assembled into a data cube, on the right.

line-free channels to obtain a continuum map, subtracting the continuum map from all the spectral channels containing line signal, and cleaning. If in the uv data then the steps are fitting a smooth (usually linear) function to the data in the line-free channels, subtracting that function from the uv data containing line signal, imaging, and cleaning. In this case the imaging and cleaning may be done simultaneously.

8.3. Data Analysis

Once the cleaned, continuum-subtracted data cube is made data analysis starts. The exact steps are determined by the goals of the research. There are more and more options in data analysis as computers become faster, less expensive, and capable of holding more data. Some examples are given in Lecture 12.

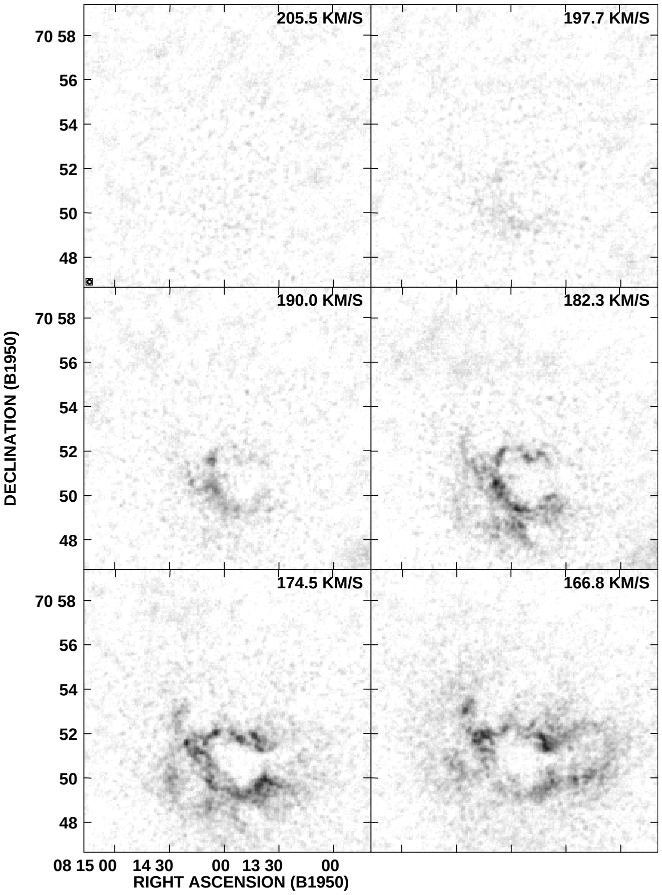


Figure 11–9. An example of a data cube of HI in emission. The nearby dwarf galaxy Holmberg II was observed with the B, C, and D configurations of the VLA by Puche et al (1992). Rotation and the filamentary nature of the HI are easily seen in the cube.

9. Examples of Spectral-Line Data Sets

9.1. HI Emission

Figure 11–9 shows the channels from a imaged, cleaned, and continuum-subtracted cube of HI emission from the nearby dwarf galaxy Holmberg II observed with the B, C, and D configurations of the VLA by Puche et al (1992). The contrast of the grey scale was chosen so that the figures are dark where the emission is strong, the strongest emission is about 20 mJy beam^{-1}. The VLA correlator was used in 2AD mode, channel width and separation were both about 2.5 km s^{-1}. Every third channel is shown to limit the size of the figure.

These observations were made to study the overall organization of the neutral ISM in Holmberg II. The motion of the HI signal from the bottom to the top in succeeding frames demonstrates the rotation of this galaxy and shows that the kinematic major axis is oriented approximately north-south. The eleven arcsecond synthesized beam is shown in the lower left corner of the first channel,

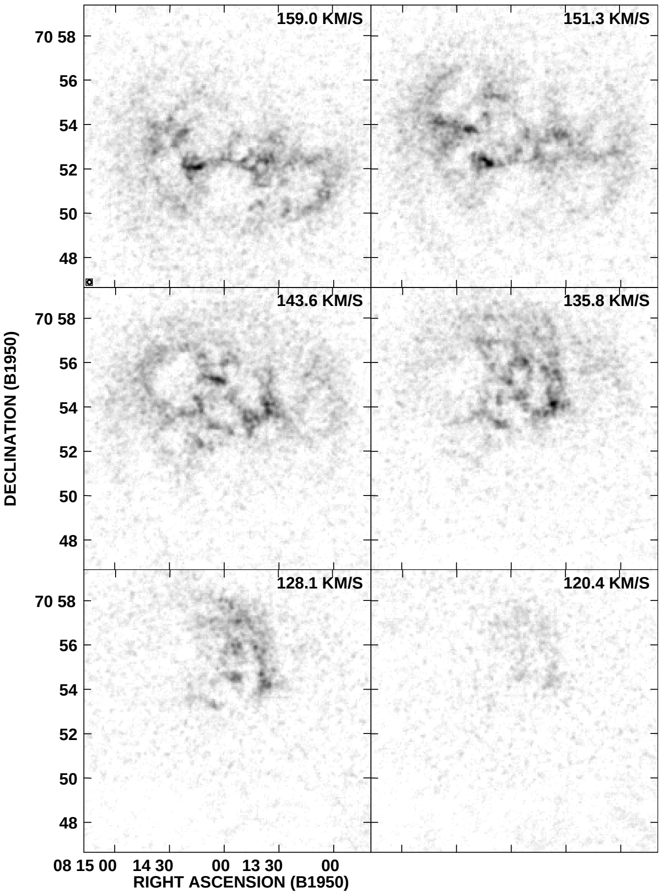


Figure 11–9. continued

it is much smaller than the diameter of Holmberg II. At this resolution holes, shells, and filaments in the HI distribution are obvious in the cube. Regions of overlapping emission in successive panels, which are separated by about 7 km s^{-1} in velocity, show that the velocity dispersion of the HI must be relatively large, and probably superthermal. Multiply-peaked spectra in the shells surrounding the holes suggest that the shells are expanding, possibly driven by the energetic events which formed the holes.

9.2. HI Absorption

Figure 11–10 shows the channels from a imaged, cleaned, continuum-subtracted cube of HI absorption against the nuclear radio continuum source of the active galaxy NGC 3894 observed with the VLBA by Peck & Taylor (1998). Their 9 × 6 milli-arcsecond synthesized beam is shown in the first panel. The deepest absorption feature is 35.9 mJy beam^{-1}, the contour separation is ten percent of that value. The contours are actually negative because the continuum has been subtracted. The nuclear source is only slightly larger than the synthesized

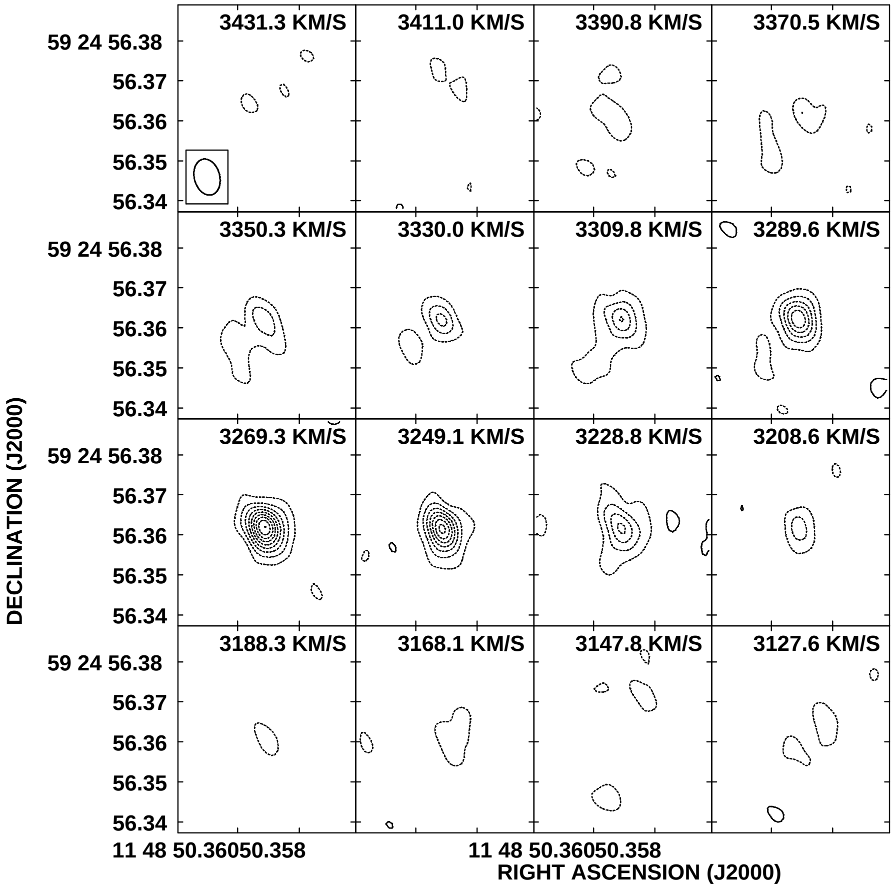


Figure 11–10. An example of a data cube of HI in absorption. Absorption against the nuclear continuum source in the active galaxy NGC 3894 was observed with the VLBA by Peck & Taylor (1998). This also serves as an example of displaying a data cube using contours. The contour interval is ten percent of the deepest absorption.

beam. The observed velocity channels were separated by about 7 km s^{-1}, every third channel is displayed to limit the size of the figure.

These observations were made in the hope of seeing a disk or torus around the central engine of this active galaxy. The synthesized beam just begins to resolve the structure of the absorbing gas, giving a hint of rotation. The great width of the HI absorption signal is obvious in the cube. Because the absorbing gas is not spatially resolved the velocity width is the convolution of the rotation and the thermal width. When spatially resolved HI absorption usually has a very narrow velocity width because the absorbing gas is cold.

It is worth pointing out that detecting absorption requires a background continuum source. The angular extent of the continuum source limits the size of the region over which absorption can be detected. It is very likely that the HI is much more extended than this cube indicates, but that lack of additional background sources makes that gas undetectable.

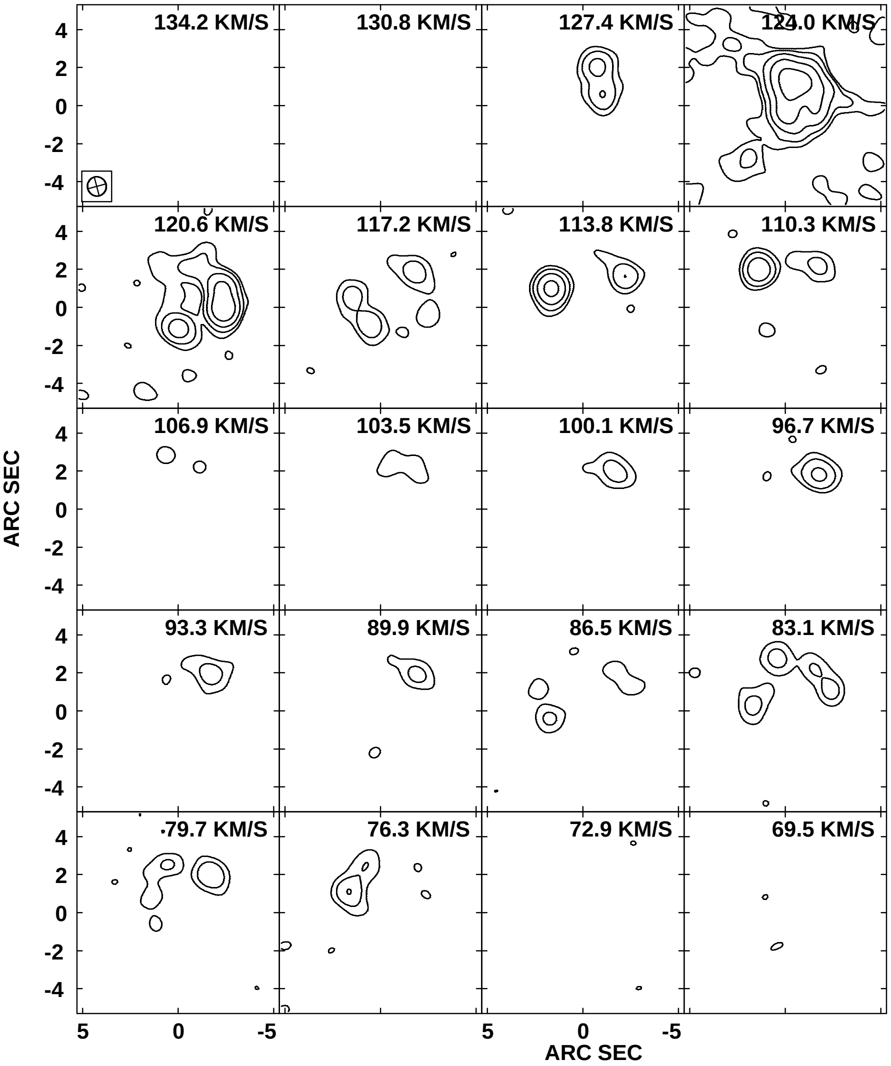


Figure 11–11. An example of a data cube of OH maser emission at 1612 MHz. The maser AFGL 2343 surrounds the GIa star HD179821, which is a candidate protoplanetary nebula. The outflow of the star's atmosphere causes the apparent change in velocity.

9.3. An OH Maser

Figure 11–11 shows a cleaned, continuum-subtracted cube of OH maser emission at 1612 MHz of the object AFGL 2343 (from the Air Force Geophysical Laboratory catalog) which surrounds the GIa star HD179821. These observations were made with the A configuration of the VLA by Claussen (1993). The maser spots are clumped about the central star, which is at position (0,0). This object is a candidate protoplanetary nebula. The channel separation was 1.13 km s^{-1}, every third channel is displayed to fit the cube on one page. Velocities are given in the LSR system, which is customary for objects within the Milky Way.

The one arcsecond synthesized beam, shown in the lower left corner of the first channel, resolves the larger maser spots but not the smaller ones. The maximum signal is 13.5 Jy beam^{-1}, and the background rms signal is 21 mJy beam^{-1}. The contours are 0.07, 0.21, 0.63, 5.67, and 17.01 Jy beam^{-1}. Note that this signal is much stronger than that of the extragalactic HI in Figures11–9 and 11–10. This is because the maser is intrinsically bright due to the amplification inherent in masers and because AFGL 2343 is galactic rather than extragalactic.

The expansion of the maser-emitting gas causes the emission to cover a range of velocities, so the peak signal moves from channel to channel. Consider an expanding sphere as a very simple model of the nebula. The brightest maser signal is in the channels of highest velocity. In the model the highest velocity occurs where the nebula is moving most rapidly away from the Sun, which is on the back of the nebula. There is no corresponding emission peak at the lowest velocities, so there is not a corresponding emission region at the front of the nebula. The combination of a three-dimensional model with the observations allows rough placement of the emission in three-dimensional space. More refined models of masers include the effects of turbulence in the nebula.

10. Conclusion

This lecture has concentrated on the information which must be gathered to plan observing and fill out the proposal form. The ultimate goal of this exercise is to advance science, which may have gotten lost in the details of the lecture. I wish to conclude by addressing the goal.

Start with a science goal and an object, or set of objects, which can be observed to meet that goal. Develop concise statements of why the goal is important and why the object is an excellent choice for observation. These statements will be invaluable for the written part of the proposal, without them it will be difficult to make a good case for observing.

Use the information in this and other lectures to choose telescopes, array configurations, correlator mode, channel size, and integration times on the object and calibrators which will give the best information for meeting your goal. This process may be iterative, in the sense that a particular telescope or correlator may not do exactly what you want, forcing you to reformulate your goal until it can be met with a real telescope. In the written proposal show how your choices are relevant to your goal.

Have a clear understanding of how the data will be reduced once the observations have been made. This is particularly important if you plan unusual or risky reduction steps. Also have a clear understanding of how the data will be analyzed and if you need to write new analysis routines.

In my experience the staff scientists at observatories are generous in helping visitors, including those who have not yet visited but are only writing their first proposal to do so. The VLA and VLBA Observational Status Summaries, available on the NRAO home page, list the scientists in Socorro and the topics on which they can provide help. Similar information is available from the other NRAO sites and other observatories. You may have benefited from their help already, many of these same people *volunteered* to be lecturers in this summer school.

Acknowledgments. I wish to thank the students in Physics 566 at New Mexico Tech, who suffered through initial versions of this material, and the students in the summer school, whose questions help to improve this written version of the lecture. I also wish to thank the observers who shared their data: Alison Peck and Greg Taylor, who made available their observations of NGC 3984, Mark Claussen, who made available his observations of AFGL 2343, and Daniel Puche, Elias Brinks, and Jean-Rene Roy, collaborators on the Holmberg II project. I wish to thank John Kormendy, Rick Fisher, and Dave Hogg, all of whom helped me to understand the formulation of good science goals, a sub-theme in this lecture. Finally, thanks to Greg Taylor, Chris Carilli, and Rick Perley for inviting me to be a lecturer. Support for image processing from the New Mexico Space Grant Consortium is gratefully acknowledged.

References

Appenzeller, I. 1997 *Reports on Astronomy*: Transactions of the International Astronomical Union, Volume XXIIIA (Reports 1996) (Dordrecht: Kluwer).

Balser, D. S., Bania, T. M., Brockway, C. J., Rood, R. T., & Wilson, T. L. 1994, *ApJ*, 430, 667.

Bania, T. M., Balser, D. S., Rood, R. T., Wilson, T. L., & Wilson, T. J. 1997, *ApJS*, 113, 353.

Burton, W. B. 1988, in *Galactic and Extragalactic Radio Astronomy*, second edition, eds. G. L. Verschuur & K. I. Kellerman (New York: Springer Verlag) 295–358.

Claussen, M. J. 1993, in *Astrophysical Masers*, eds. A. W. Clegg & G. E. Nedoluha (Berlin: Springer-Verlag) 353–356.

Ewen, H. I. & Purcell, E. M. 1951, *Nature*, 168, 356.

Field, G. B. 1959, *ApJ*, 129, 536.

Giovanelli, R. & Haynes, M. P. 1988, in *Galactic and Extragalactic Radio Astronomy*, second edition, eds. G. L. Verschuur & K. I. Kellerman (New York: Springer-Verlag) 522–562.

Gordon, M. A. 1988, in *Galactic and Extragalactic Radio Astronomy*, second edition, eds. G. L. Verschuur & K. I. Kellerman (New York: Springer-Verlag) 37–94.

Heiles, C., McCullough, P. R., & Glassgold, A. E. 1993, *ApJS*, 89, 271.

HöguInd, B. & Mezger, P. G. 1965, *Science*, 150, 339.

Jorgensen, U. G. 1993, *Molecules in the Stellar Environment* (New York: Springer-Verlag).

Kerr, F. J. 1968, in Stars and Stellar Systems, volume VII, *Nebulae and Interstellar Matter* eds. B. M. Middlehurst & L. H. Aller (Chicago: University of Chicago) 575–622.

Madore, B. F. & Tully, R. B. 1986, *Galaxy Distances and Deviations from Universal Expansion* (Dordrecht: Reidel).

Mezger, P. G. & Höglund, B. 1967, *ApJ*, 147, 490.

Miao, Y., Snyder, L. E., Kuan, Y.-J., & Lovas, F. J. 1994, *BAAS*, 26, 906.

Muller, C. A. & Oort, J. H. 1951, *Nature*, 168, 357.

Nenner, I. 1993, *Molecules and Grains in Space*, AIP Conference Proceedings 312 (New York: American Institute of Physics).

Pawsey, J. L. 1951, *Nature*, 168, 358.

Peck, A. B. & Taylor, G. B. 1998, *ApJ*, 502, 23.

Puche, D., Westpfahl, D. J., Brinks, E. & Roy, J.-R. 1992, *AJ*, 103, 1841.

Reid, M. J. & Moran, J. M. 1988, in *Galactic and Extragalactic Radio Astronomy*, second edition, eds. G. L. Verschuur & K. I. Kellerman (New York: Springer-Verlag) 255–294.

Seidelman, P. K. 1992, *Explanatory Supplement to the Astronomical Almanac* (Mill Valley: University Science Books).

Shklovsky, I. S. 1949, *Astronomicheskii Zhurnal*, 26, 10.

Sullivan, W. T. 1982, *Classics in Radio Astronomy* (Boston: D. Reidel).

Thompson, A. R., Moran, J. M., & Swenson, G. W. 1991, *Interferometry and Synthesis in Radio Astronomy* (Malabar: Kreiger).

Tongue, T. D. & Westpfahl, D. J. 1995, *AJ*, 109, 2462.

Turner, B. E. 1974, in *Galactic and Extragalactic Radio Astronomy* eds. G. L. Verschuur & K. I. Kellerman (New York: Springer-Verlag) 199–255.

Turner, B. E. 1988, in *Galactic and Extragalactic Radio Astronomy*, second edition, eds. G. L. Verschuur & K. I. Kellerman (New York: Springer-Verlag) 154–199.

Turner, B. E. & Ziurys, L. M. 1988, in *Galactic and Extragalactic Radio Astronomy*, second edition, eds. G. L. Verschuur & K. I. Kellerman (New York: Springer-Verlag) 200–254.

van de Hulst, H. C. 1945, *Nederlandsch Tijdschrift voor Natuurkunde*, 11, 210.

Verschuur, G. L. 1974, in *Galactic and Extragalactic Radio Astronomy* eds. G. L. Verschuur & K. I. Kellerman (New York: Springer-Verlag) 27–50.

Weinreb, S., Barrett, A. H., Meeks, M. L., & Henry, J. C. 1963, *Nature*, 200, 829.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

12. Spectral Line Observing II: Calibration and Analysis

M.P. Rupen
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. This lecture discusses the unique aspects of calibrating, imaging, and analyzing multi-frequency aperture synthesis data. Special attention is given to bandpass calibration, continuum subtraction, and some of the more common and generic analysis problems.

1. Introduction

The reduction and analysis of multifrequency data differ in several ways from those of continuum observations. In part these differences are based on practical considerations – why perform the same operations for every spectral channel, if there are short cuts available? – but there are substantive differences as well, mostly to do with avoiding systematic effects in comparing or integrating the data from multiple channels. This lecture discusses the problems unique to spectral line data and the most common methods used to address them, proceeding from calibration through data analysis. The emphasis is on basic ideas and general approaches rather than specific reduction packages or VLA techniques. As always the observer should consider her own instrument and science when deciding how best to reduce her own data. The purpose of this lecture is to provide an informed basis for such decisions.

2. Calibrating the Bandpass

The goal of calibration is to take imperfect data, corrupted by atmospheric and instrumental effects, and reconstruct from them (insofar as possible) the data that would have been taken by a perfect telescope unaffected by the atmosphere. In continuum calibration, as discussed in Lecture 5, one tries to account for temporal and spatial variations in the gain. For spectral line data one must worry about frequency variations as well. The relative gain of an antenna or baseline as a function of frequency (channel) is called the *bandpass*.[1]

2.1. Why is the Bandpass Not Flat?

In an ideal interferometer under a non-dispersive sky, the instrumental response would be the same at all frequencies, i.e. the bandpass would be flat with amplitude 1.0 and phase 0.0. In reality a variety of effects ensure that this is not the case:

- **The atmosphere** introduces variations which are important only over much wider bandwidths than current interferometers employ (though the proposed VLA Upgrade, and possibly the Millimeter Array, may change this, especially at 40–50 GHz).

[1] Sometimes called the *passband*.

- **Standing waves** (stable modes of oscillation) between the feed and the subreflector are excited primarily by receiver noise leaking out from the feed; there is also a smaller component which is excited by external sources, e.g. the Sun, the ground, and the source being observed. The characteristic frequency scale is half the focal length. This is the dominant source of bandpass ripples for single-dish telescopes[2], but is generally negligible for interferometers, because the receiver noise does not correlate from one telescope to the next, and because the signals from external sources are enormously reduced by delay beam smearing (see §5. below).

- **Front-end systems:** the filters and amplifiers in the receiver box, and to a lesser extent the feeds, do not have a flat response with frequency, and this will be reflected in the antenna gains.

- The **IF transmission system** from the antenna to the correlator may also introduce some frequency dependence. Generally these are delay errors (phase slopes with frequency), introduced by dispersion in the delay lines and the like, and since they are fairly stable can be removed through occasional monitoring measurements. The VLA has in addition the particularly nasty problem of standing waves within the waveguide, which lead to a $\sim 3\,\mathrm{MHz}$ ripple which shifts in phase as the ambient temperature changes (Carilli 1991; Lilie 1994).

- The **back-end filters** located near the correlator, like the front-end filters in the receiver box, also have some frequency dependence, beyond the intended sharp cutoff.

- Finally, the **correlator** itself generates some frequency variation. Most modern interferometers use digital correlators, which approximate the incoming digital signal using a few (generally one or two) bit sampling (see Lecture 4). This quantization introduces a significant bias when the resulting correlation coefficient is large, as will generally be the case for strong continuum signals at short lags, and for strong lines throughout. This bias can be calculated and corrected (the Van Vleck correction: see Lecture 4, and pp. 216 ff. of Thompson, Moran, & Swenson 1991 (TMS)), though this is only just now being done at the VLA. More important is the Gibbs ringing introduced by the finite time/frequency sampling of the signal, which generates frequency ripples spreading out from any sharp edges in the signal. This is discussed at length in §3. below.

This is a rather daunting list, but fortunately most of these effects (for a well-designed system) do not change much with time, requiring only occasional calibration. Further, most of the frequency dependence is antenna-based, making the bandpass an obvious candidate for frequency-dependent self-calibration. Exceptions include errors arising in the correlator, which apparently dominate the

[2] Note that bandpass ripples are usually called 'baseline ripples' in the single-dish context, a usage avoided for obvious reasons by interferometrists!

residual bandpass at BIMA, and mis-matches in the delay line at those interferometers (notably BIMA and OVRO) which use different delay lines for every baseline.

2.2. Splitting the Time and Frequency Dependence of the Gain

The stable part of the frequency response can be removed easily through occasional calibration measurements. Thus the delays are measured at most interferometers on a regular basis, and removed by the on-line system before the observer sees the data. These corrections cannot be measured perfectly, at all frequencies, or with all possible correlator/filter/IF combinations, so there remain both residual, relatively constant gain errors, and those which vary significantly with time. The former require only occasional (e.g., daily) calibration with the actual observational setup, while the latter are more pernicious, with the VLA's "3 MHz ripple" requiring calibration as often as once every hour or less. Since the bandpass often varies significantly with frequency, it must be determined separately for each observed frequency, for instance when observing H I in galaxies with different recession velocities.

Measuring the bandpass can be time consuming, since it requires reasonable signal-to-noise ratios in each narrow channel. Thus we have a spectral response that is difficult to determine but which changes only on time scales of hours to days, superposed on an overall response which is much easier to measure but which can vary (due to electronic drift and atmospheric changes) on much shorter time scales. It is therefore advantageous to divide the overall calibration into two semi-independent steps, measurement of the average response on relatively short time scales, followed by measurement of the spectral variation about that average on time scales as long as the instrument permits:

$$\mathcal{G}_{ij}(\nu,t) = \mathcal{G}'_{ij}(t)\,\mathcal{B}_{ij}(\nu,t)\,, \tag{12–1}$$

where $\mathcal{G}_{ij}(\nu,t)$ is the total gain on baseline $i-j$; $\mathcal{G}'_{ij}(t)$ is the time-variable continuum gain; and $\mathcal{B}_{ij}(\nu,t)$ is the frequency-dependent part of the gain, which may vary slowly with time. Note that all these gains are complex numbers. The assumption is that the frequency dependence is sufficiently mild that the signal does not decorrelate when averaging across the band; if this is true one can average the raw data in frequency to gain sensitivity when measuring time variability[3]. For VLBI observations this is not always the case, and one is forced to solve for both the time and a linear approximation to the bandpass (the delay) at the same time (see Lecture 22).

If the overall gain and the bandpass can indeed be solved for independently, the order in which the two steps are done is largely a matter of taste, except in certain unusual circumstances. At the VLA we tend to remove the time variability before doing the bandpass solution, but actually they are done independently and the order is irrelevant. The main benefit in looking at the

[3] The orthogonal assumption, that the time dependence is mild and can be ignored when determining a first-order bandpass, is less often true. It is also seldom necessary, since dividing by the pseudo-continuum (see below) will remove any overall time dependence automatically.

frequency-averaged data first is that the data set is significantly smaller, which makes overall quality checks and data flagging easier.

2.3. Determining the Bandpass

The bandpass is determined by observing a source of known spectral shape and comparing the apparent to the intrinsic spectrum. The simplest approach is to observe an astronomical source which has a constant flux across the observed bandwidth (zero spectral index: $\alpha = 0$, $S_\nu \propto \nu^\alpha$). The bandpass calibrator need not be a point source, so long as its apparent structure does not change over the band: since we are interested in the *relative* gain as a function of frequency, we can remove the source structure by dividing the visibilities at each frequency by those of the continuum, leaving a spectrum for each baseline which depends only on the bandpass. (This step is not necessary if the calibrator is intrinsically point-like; in that case one can simply read the bandpass off from the data directly, modulo a scaling factor for the flux density.) This requires some estimate of the continuum visibilities, which may be derived from the line data themselves (as at the VLA) or from simultaneous broad-band observations (as at various of the millimeter interferometers). Usually the normal continuum calibration (solving for the overall gain variations as a function of time) is done using these continuum data as well.

The frequency spectrum of the time-averaged, continuum-divided visibilities gives a direct estimate of the bandpass for each baseline. If the signal-to-noise ratio (SNR) is high enough, this may be applied directly to the on-source data. This happy circumstance is seldom achieved in practice, and in any case most of the bandpass variations (as with gain variations in general) are intrinsically associated with individual antennas, not individual baselines. So the bandpass is usually approximated as the product of antenna-based bandpasses

$$\mathcal{B}_{ij}(\nu, t) \approx b_i(\nu, t)\, b_j^*(\nu, t) \tag{12–2}$$

and a channel-by-channel self-calibration, using the continuum-divided visibilities, gives the antenna-based bandpasses $b_i(\nu, t)$ (Figure 12–1). For large numbers of antennas this improves the accuracy of the bandpasses considerably, as one uses $N(N-1)/2$ baseline spectra to derive N complex gains. The current generation of millimeter interferometers originally had few enough antennas that self-calibration did not help much, so they often rely on baseline-based bandpasses.

A channel-by-channel calibration, even if antenna based, requires a reasonably strong source, and perhaps a fairly long integration (see Lecture 10 for a discussion of noise in self-calibration). Since such sources may not be available, particularly at high frequencies, and observing time is always at a premium, various approaches may be taken to improve the SNR of the bandpass.

- **Autocorrelation bandpasses:** One obvious source of stronger signal is the sky (atmosphere) itself, which is much brighter than most sources at centimeter and millimeter wavelengths. Autocorrelation spectra may thus be used to determine the amplitude part of the antenna-based bandpasses quickly and continuously. Not only does this track even short time scale variations, the autocorrelations come "for free," requiring no special observations. However, this approach has been dropped from most modern

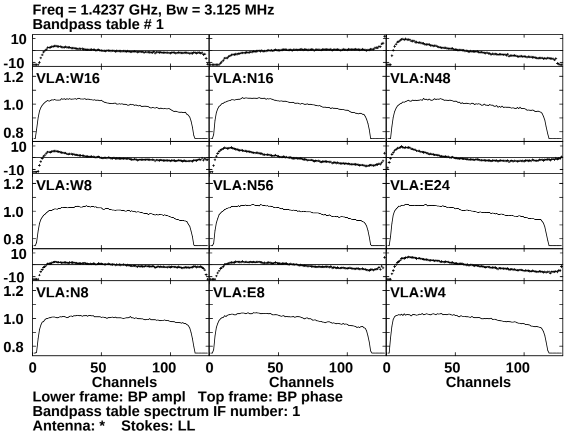


Figure 12–1. Antenna-based bandpasses for a VLA H I absorption experiment with 127 channels covering 3.125 MHz. Before solving for the bandpass the data were divided by a pseudo-continuum to remove the source structure. This so-called "channel 0" was formed by vector-averaging the inner three-quarters of the frequency channels. Hence the average phase in that range is near 0, and the average amplitude near 1. Note the residual delay errors (phase slopes).

interferometers (the Plateau de Bure instrument remains an interesting exception), for two reasons: first, the phases cannot be calibrated in this fashion; and second, the autocorrelations are corrupted by problems that interferometry minimizes or eliminates entirely (standing waves, radio frequency interference, etc.). The autocorrelations are also not vulnerable to Gibbs ringing at the lower (base) band edge, and so cannot be used to calibrate out that effect (see §3.).

- **Injected noise:** Another approach is to insert a noise source at the antennas and measure the resulting cross-correlations. One can determine both the amplitude and the phase gain, apart from electronic effects ahead of the insertion point. Standing waves and the like are avoided by injecting the signal after the feed, but one must be sure the noise source itself has no (or known) frequency structure, and there may be some residual bandpass errors ahead of the injection point. Also of course this takes some time from the observing program, but since the noise source can be reasonably strong the time lost is minimal.

- **Cutting down the number of parameters:** An alternative to increasing the signal is to decrease the number of variables which must be solved for. Apart from self-calibration, most methods rely on the assumption that the bandpass is intrinsically rather smooth, which will be true when using the narrow channels (high frequency resolution) that give the lowest initial SNR (but see the section on Gibbs ringing, below). Possibilities include smoothing the data or the derived bandpasses in frequency (e.g., the AIPS task `BPASS`) or fitting them with some functional form (generally a low-order polynomial; e.g. the AIPS task `CPASS`).

In practice many interferometers use hybrid schemes. Millimeter observatories in particular are plagued by a lack of strong sources, and either maintain a standard set of default bandpasses for many possible observational setups, or employ a noise source to provide sufficient power to measure the bandpass (see e.g. Wright 1991, 1994). At the Plateau de Bure the noise source is used only to determine the phase, while the autocorrelation spectra give the amplitude. A final, corrective bandpass is determined using an astronomical calibrator to remove residual delay errors due to electronic components before the noise insertion point, and any structure in the noise source itself. However, weaker objects can now be used as calibrators, because one is determining only a few additional parameters of the bandpass (generally, a linear fit to the amplitude and phase as a function of frequency), as the noise source (sometimes called IF) bandpass has already taken out the small-scale, channel-to-channel variations. Another interesting variant is the "pulse cal" system used at the VLBA, in which several pure tones are injected and the delay and rate measured by fitting low-order polynomials through the phases at those frequencies; a standard astronomical bandpass may then be used to remove the residual errors.

Figure 12–1 shows a typical set of antenna-based bandpasses for the VLA. The visibilities from which these were derived were normalized by the pseudo-continuum "channel 0" data, the vector average[4] of the visibilities over the inner 3/4 of the band. Hence by construction the bandpass over those inner channels has mean phase 0 and mean amplitude 1. The fall-off at low and high frequencies is characteristic of the filter shapes; note that this decline does not necessarily imply high noise levels or bad data, since the instrument is linear over a large range. The very edge channels are indeed noisier than those in the middle, but VLA scientists are still debating why this is so. In general narrower bandwidths will have flatter frequency responses, because the elements of the system have fairly wide characteristic bandwidths — it is simply difficult to create bandpass structure on scales of, say, a few kHz.

2.4. Checking and Using the Bandpass

The bandpass once derived is a powerful tool for finding problems with spectral line data; so much so that some observers derive bandpasses from their line sources as well as their calibrators, simply to check the quality of their data. This is because most important errors are antenna-based, so checking antenna-based bandpasses is often the simplest and most appropriate way to find them. Radio frequency interference (RFI), for instance, is often associated with a certain antenna or set of antennas, because the RFI is generated near that antenna. Narrow RFI signals will show up as large spikes or dips in the bandpass for that antenna, which can then be flagged. More subtle effects can be found by comparing the bandpasses for different antennas and as a function of time;

[4] The *vector average* of a series of complex numbers follows the usual mathematical definition of a complex average, and may be obtained by separately averaging the real and imaginary parts. This contrasts with the *scalar average*, in which the amplitudes and phases are averaged instead. The scalar average is sometimes useful in the presence of large phase errors (especially in uncalibrated data) where the phase jitter would give a vector average of zero, while the mean amplitude might still be useful, e.g. in estimating a flux density. The ratio of the amplitudes of the vector and the scalar averages provides a simple estimate of the coherence of the signal.

in general most bandpasses should look fairly similar (and have similar noise levels), and there should be no abrupt changes with time. Errors in the channel-dependent self-calibration can also be useful diagnostics. In careful reductions one may iterate through the bandpass/flagging loop several times, to derive the best possible bandpass while removing as much "problem data" as possible.

The accuracy of the bandpass itself is important, particularly when there is strong continuum emission. An error in the bandpass can produce spurious line emission or absorption on top of continuum features, precisely where real emission is likely to be; even without continuum emission, phase errors can cause a stationary source to appear to move from channel to channel. On a more fundamental level, noise in the bandpass creates (systematic) noise in the channel maps, again particularly when there are strong signals. This is all precisely analogous to the effects of standard (continuum) gain errors (Lectures 5, 15).

2.5. Complications and Subtleties

- **Bandpass interpolation:** For most interferometers and in most observations, the bandpass is stable enough that an occasional determination suffices, and the precise method used to apply that determination to observations taken somewhat earlier or later does not really matter. Unfortunately at the VLA this is not true. The infamous "3 MHz ripple" (Carilli 1991) drifts in frequency as the temperature changes, introducing significant variability on time scales of a few hours. For detailed work one must both calibrate often and interpolate intelligently, with several different methods being implemented in the AIPS option `DOBAND` and the AIPS task `BPSMO`. An instructive example is given in van Gorkom (1993).

- **Source structure:** Most bandpass determinations rely on the calibrator not changing with frequency, i.e. having an intrinsically flat spectrum and a structure that does not change (as sampled by the interferometer) over the observed range of frequencies. There are two major cases where these assumptions fail: wide bandwidths and Galactic observations. Over wide bandwidths, the calibrator's spectral index will introduce flux variations, while the changing (u, v) coverage will sample significantly different Fourier components of the source structure; this is the basis of multi-frequency synthesis (see Lecture 21) but in the current context is a nuisance. One must either build up a frequency-dependent model of the calibrator, or break the observation up into smaller sub-bands. In Galactic observations, e.g. H I absorption studies, it is likely that the calibrator itself will have line emission or (more frequently) absorption at the observed velocity. In this case there are three options: find a calibrator without a line at this frequency, which means finding one at very different Galactic coordinates; observe the calibrator at different frequencies than the source, and interpolate the bandpass in frequency; or flag the affected calibrator channels and interpolate the bandpass across the resulting frequency 'hole.'

- **Channel-dependent noise:** In general the bandpass determination itself will be noisy, this noise will depend on both frequency and baseline, and these errors in the applied bandpass will produce corresponding errors in the data for the source one is actually interested in. Most obviously this occurs when using an extended calibrator, because the long spacings see less flux than the short ones, and hence yield less accurate estimates of the bandpass. This has two effects: first, the varying noise as a function of baseline should be taken into account when doing the channel-based self-calibration; and second, the noise in the bandpass should be propagated through to channel-dependent weights for the source's (u, v) data (see Lecture 7). These subtleties are only now beginning to appear in real data reduction packages; they are in any case only important for delicate experiments.

- **Time-variable frequencies:** As discussed in §4.4., many line observations are carried out at constant velocity, whereas the bandpass is inherently a function of observed (topocentric) frequency.

- **Polarization bandpasses (frequency-dependent D-terms):** The polarization D-terms, which measure the leakage between the two orthogonal polarizations, may have significant frequency dependences (e.g. at the VLBA; see Lecture 24). These must be corrected by calibrating the polarization bandpass. Currently only `aips++` offers software to handle this in a straightforward fashion. The ionospheric Faraday rotation might also be considered a bandpass effect, though it is normally removed as a separate step in the data reduction.

2.6. The Bandpass and Continuum Observations

Continuum observers generally just measure the time dependence of the overall gain, and ignore variations with frequency. But the entire purpose of bandpass calibration is to remove those variations — why do we have to worry about this for spectral line and not for continuum observations? There are two answers to this. First, bandpass calibration is more important for spectral line observations because we are interested in the detailed comparison of images made at different frequencies, and because we do not wish to confuse apparent changes in continuum emission caused by gain errors with real line emission or absorption. Second, there is a price to simplicity, even for continuum observations. Uncalibrated bandpass effects are among the dominant errors in continuum images. Without knowing the bandpass, or equivalently the filter shape, one does not know the effective central frequency of the observed band; but that frequency sets the scale for the baseline lengths and hence the image/pixel size (see the discussion of chromatic aberration, §5.), and getting it wrong leads to a radial distortion of the image. This effect can be large — at the VLA, the nominal "50 MHz" band used by default for continuum observations has an actual bandwidth of about 43 MHz, and its sensitivity-weighted central frequency is actually a few MHz from that specified by the observer. (This is one of the major effects which had to be considered in the NVSS — cf. Condon *et al.* 1998.) A more subtle but insidious effect comes from undersampling the antenna-based bandpass. If the frequency response is not flat within a single channel, and differs from antenna to antenna, the resulting errors in the visibilities (cross-correlations) do not close — they cannot be decomposed into antenna-based gains. This is because the response to a signal in this wide channel is no longer simply the product of the two antenna's gains, but instead an integral of that product over frequency. The resulting non-closing errors cannot be fixed by self-calibration, and the only way fully to avoid them is to take continuum data using spectral channels narrow enough to track the frequency structure of the bandpass.

3. Gibbs ringing

In almost all synthesis telescopes, the spectral line capability is provided by a digital cross-correlation spectrometer. Although this avoids many errors involved in an analog system, it introduces an additional effect, the *Gibbs phenomenon.* This is the 'ringing' which occurs near sharp changes in the frequency spectrum because of the truncation of the lag spectrum, and corresponds to the residual error when one tries to model a sharp transition with a finite number of Fourier components. This effect is most obvious for a very narrow spectral line, which with finite resolution appears as a delta function. If the filter which sets the total bandwidth has a sharp upper- or lower-frequency edge, this will also cause ringing. Less obviously, there will also be a sharp transition at baseband, i.e. at zero frequency. This is easiest to see in the case of a real correlator like that employed at the VLA. Such correlators generate a frequency spectrum which is Hermitian, i.e. give the complex conjugate of the "real" spectrum at negative frequencies (Figure 12–2). This gives a phase discontinuity at $\nu = 0$ if the observed visibility phase is non-zero, which it usually is for uncalibrated data. The resulting ripple will change with the observed phase, i.e. will vary with position on the image and with the instrumental phase. The effect will be correctly removed by the bandpass calibration only at the position of the bandpass calibrator (if that is a point source, and if the instrumental phase has not changed between bandpass calibration and source observation).

Quantitatively, the truncation of the observed lag spectrum corresponds to multiplying the true lag spectrum by a box function. The effect in frequency space then is the Fourier transform of this, i.e. the convolution of the true spectrum with a sinc function ($\sin x/x$), with the nulls of that function spaced by the channel separation. Hence XF correlators, which multiply the antenna-based signals and then Fourier transform, produce a sinc response; while FX correlators, which Fourier transform antenna-based signals and then multiply, give the more benign sinc^2 (see Lecture 4 for a fuller discussion). A monochromatic signal centered exactly on a channel will appear as a simple spike in the output spectrum; however, spectral lines are seldom so narrow and so perfectly placed, and in general one has to deal with the 22% spectral sidelobes of the sinc function. The entire band can be swamped by uncalibrate-able frequency oscillations; as noted above this may occur even without the presence of a sharp line. There are three possible remedies (for concreteness hereafter I assume the XF design, as used at most connected-element interferometers, but *not* at the VLBA):

(1) **Observe with a large number of channels**, i.e. increase the length of the lag spectrum. The channels most strongly affected by ringing can simply be discarded. The oscillations reduce to about 2% around channel 20, so with a large enough correlator this is not unreasonable. For ringing

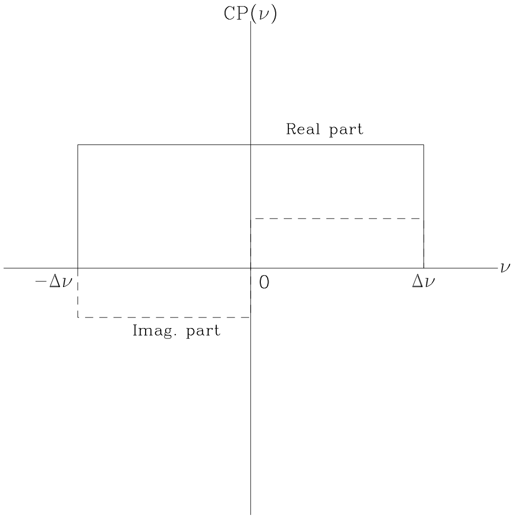


Figure 12–2. The sharp edge at baseband. Since the lag spectrum is real the frequency spectrum is Hermitian, and any non-zero phase corresponds to a sharp change in the imaginary part across the origin. (CP is the correlated- or cross-power.)

of sharp spectral features this throws away what are probably the most interesting data, but for band-edge problems this is a practical alternative.

(2) **Smooth the data in frequency**, i.e. taper the sharp edge of the lag spectrum. Basically one trades frequency resolution for sidelobe suppression, a familiar tradeoff in the image plane (see Lecture 7). As with weighting in the (u, v) plane, there is a wide variety of smoothing kernels from which to choose depending on what is desired: minimization of distant or near-in spectral sidelobes, maximization of spectral resolution or sensitivity, etc. Harris (1978) provides a useful summary. Unfortunately this flexibility is seldom exploited or available within current software packages, and Hanning smoothing is by far the most common (and usually the only) option. Its main selling point is its simplicity: the smoothed spectrum is

$$S_h(\nu_i) = \frac{(S(\nu_{i-1}) + 2S(\nu_i) + S(\nu_{i+1}))}{4} \qquad (12\text{–}3)$$

where i labels the i^{th} channel. The full-width at half maximum (FWHM) of the sinc function is $1.2\Delta\nu$, where $\Delta\nu$ is the channel spacing. After Hanning smoothing this rises to $2.0\Delta\nu$, but the sidelobes are reduced below 3% (Figure 12–3). The noise-equivalent bandwidth also increases from $1.0\Delta\nu$ for the sinc function, to $2.0\Delta\nu$ after Hanning smoothing.

(3) **Account for the Gibbs phenomenon during the data reduction.** In a few special cases — e.g. a line source with continuum emission whose structure matches that of the bandpass calibrator — the Gibbs ringing

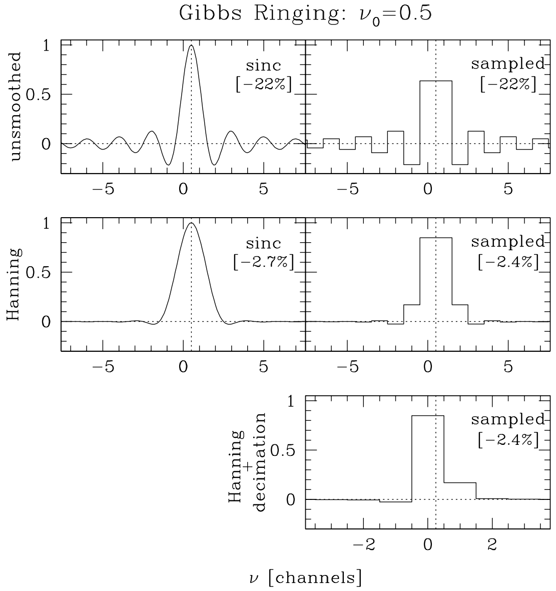


Figure 12–3. The spectral response of a finite-size XF correlator, before and after Hanning smoothing and decimation. This figure illustrates the worst case, an infinitely narrow spectral line located halfway between channels 0 and 1. The amplitude has been normalized to unity, i.e. the integral under each of these curves is 1.0. The left-hand panels show the sinc function itself, while the right-hand panels show what would actually be observed. The peak negative sidelobe of both the sinc and the observed function is −22% before smoothing; Hanning smoothing reduces this by roughly a factor of 10, as given in the figure, at the expense of a considerable broadening of the central response. At the VLA observers often discard half the channels after Hanning smoothing, giving a spectrum like that in the bottom panel; this makes it more difficult to reconstruct the true profile, although it does make the data sets smaller by a factor two.

will be calibrated out as part of the standard bandpass. If the line source has no sharp features and no continuum, one can Hanning smooth the data for the bandpass calibrator (to avoid introducing Gibbs ringing in the bandpasses) while retaining the full spectral resolution for the line source. Generally however one requires three-dimensional deconvolution and modeling software to account for spectral as well as spatial sidelobes. So far no such software is readily available in the standard synthesis packages, which tend to conceive of spectral line data as stacks of independent, continuum-like channels.

For other discussions of the Gibbs phenomenon, see TMS, pp. 236–241; van Gorkom & Ekers (1989); Wilcots, Brinks, & Higdon (1995); and Bos (1984, 1985).

4. Special calibration topics

4.1. Self-calibration

Some spectral lines can be strong enough to allow self-calibration (Lecture 10). Masers are ideal, since they can be very strong, are generally compact, even on VLBI scales, and (for H_2O and SiO in particular) can occur at frequencies where the atmosphere is not very stable. Indeed, masers have at times been used to phase-reference observations of weak continuum sources, notably in evolved stars and star-forming regions (e.g., Reid & Menten 1990; Torrelles *et al.* 1996, 1997) and in the spectacular Seyfert galaxy NGC 4258 (Herrnstein *et al.* 1997). Similarly, weak spectral lines observed in fields with strong continuum sources can benefit from self-calibration of that continuum. Even with self-calibration one must still measure the bandpass independently (except in very special circumstances; see van Gorkom *et al.* 1993), because self-calibration inherently measures either an average gain or the gain at a specific frequency. This can be enormously helpful in removing the time dependence of the atmosphere, but allowing frequency dependences into the self-calibration process will usually distort, and possibly remove, the very line signal one is trying to detect.

4.2. Strong and Ubiquitous Lines

Especially strong and/or pervasive spectral lines, e.g., masers or Galactic H I, lead to some unusual situations:

(1) *Bandpasses:* As mentioned above, one may have to switch or interpolate across frequencies when observing calibrators to avoid line contamination, and this will limit the accuracy of the bandpass, and hence the fidelity and spectral dynamic range (see §8.) of the images. Some lines are also strong enough to corrupt the autocorrelation spectra, disallowing their use for bandpass determinations.

(2) *Amplitude calibration:* Lines strong enough to be seen in single-dish spectra may be used to cross-calibrate the amplitude gains of the antennas, by scaling each autocorrelation to match a template spectrum. This is frequently done for VLBI observations of masers (e.g., Lecture 24).

(3) *System temperature corrections:* Strong lines will raise the system temperature significantly, particularly for narrow bandwidths. If the visibility data are scaled by the system temperature, as is done at many telescopes including the VLA and the VLBA, one must be careful to use the system temperatures appropriate to the observed bandwidth.

(4) *Corruption of 'continuum' data by spectral lines:* This is seldom a problem at centimeter wavelengths where lines are few, widely spaced, and generally weak, but for millimeter observers line contamination is a serious issue. See Harris *et al.* (1995) and Schilke *et al.* (1997) for some high-frequency estimates.

4.3. Joining Adjacent Frequency Bands/IFs

Life being what it is, one would occasionally like to observe a broader frequency range than can be accommodated by a single intermediate frequency (IF) band or baseband channel (BBC). The most difficult problem is observing a single very broad line, where one must ensure that the calibration errors across the IF boundary do not mimic or corrupt the true line. Cross-calibration between IFs is intrinsically more difficult than calibration within a single one, because the IFs go through different electronics, and because filters and other components tend to have their most rapidly changing (and in some cases degrading) response near band edges, precisely where one wishes to join the IFs together. How bad this problem is depends on the instrument, but in general it is wise to overlap the IFs by some amount, to allow direct checks of their mutual calibration and to ensure that band-edge effects do not wipe out the most interesting (central) portion of one's line.

4.4. Frequency, Velocity, and Doppler Tracking

Generally one wishes to observe astronomical spectra in the rest-frame of the emitting object, for example to identify the lines, or to interpret their Doppler shifts in terms of kinematics intrinsic to the source. Since the available telescopes are usually moving with respect to those astronomical sources, the observing frequency must be continually adjusted to maintain the same source velocity (see Lecture 11). This Doppler tracking is often done on-line (e.g., at the VLA), but in some cases (e.g. for the ATCA and VLBI) the observing frequency is kept constant, and the conversion to source velocity done after the fact. In this latter case a given velocity channel in the final cube must be constructed from a range of observed frequencies, and will seldom correspond exactly to any one observed frequency channel. Common approaches to the problem are to make the desired velocity channel out of a weighted sum of the input frequencies (Miriad's `line` parameter), or to do an FFT-based interpolation to the appropriate frequency (as in `aips++`, or with AIPS' `CVEL`). The latter is more accurate but also more time-consuming. Note that the baseline lengths should be measured in units of the original frequencies; this is seldom an important distinction for a single observation, but may occasionally be important when combining data taken at different times during the year. The distinction between frequency and velocity complicates bandpass determinations as well, since the bandpass is fundamentally related to the frequency response of the instrument rather than any sky velocity. Finally, since RFI tends to be fixed in frequency rather than velocity, Doppler tracking may lead to a single RFI spike corrupting several velocity channels. One should therefore remove (flag) such interference before moving to the rest frame of the astronomical source.

5. Chromatic Aberration (Bandwidth Smearing)

Radio synthesis telescopes have an inherent chromatic aberration, often referred to as bandwidth or delay beam smearing, because they form images by adjusting the phase $\Delta\phi$ rather than the arrival time $\Delta t = \Delta\phi/2\pi\nu$ of the correlated signals for each point in the image[5]. For observations of finite bandwidth this leads to

[5] Optical interferometers avoid this by inserting delay lines before down-converting the incoming frequencies; most radio interferometers down-convert to an intermediate frequency before

a radial smearing that increases linearly away from the point in the image for which the time delays were equalized (known as the *delay tracking center*). This aberration may be kept to a minimum by observing with very narrow channels, i.e. in spectral line mode, since one may then use the actual central frequencies of each channel rather than the band center frequency when constructing the images. For wide-field continuum images the maximum bandwidth per channel will be determined by the desire to avoid this delay beam smearing.

Although the use of narrow frequency channels can minimize delay beam smearing, there is still the problem (or the feature; see Lecture 21) that the (u, v) coverage itself changes with frequency. This is because the Fourier spacings to which an interferometer is sensitive scale with frequency, and implies that the point spread function (sidelobe structure) will too. This must be taken into account when analyzing the spectral line data cubes. Deconvolution reduces the effects of a changing point spread function, but even so it is better to proceed as far as possible before introducing the artifacts such non-linear algorithms can create.

5.1. Single vs. Double Sideband

Many single-dish telescopes use double sideband rather than single sideband receivers, which for the observer means that each channel corresponds to two sky frequencies (separated by an amount which depends on the details of the tuning system). This has the advantage that one can potentially observe two spectral lines at once, and double the bandwidth for continuum observations; on the other hand, such data are obviously more complicated to analyze, and a line in one sideband will have noise added in from the other as well. The problem is more fundamental for interferometers, since one has to stop the fringes in both sidebands independently. This complicates the electronics, requiring two sets of fringe offsets, although the sidebands can also be separated at the correlator output by a 90° phase-switching technique. See TMS for details. One nice feature of such systems is that the sidebands are separated explicitly, before the observer has to deal with the data; this eliminates the single-dish problems mentioned above. Millimeter interferometers, operating in a regime where multiple lines are common, bandwidth is readily available, and continuum sources are usually faint, often offer the observer the option of double sideband observations.

6. Continuum Subtraction

Continuum subtraction is generally one of the first steps in the data reduction following application of the bandpass. The reasons for subtracting the continuum are many. It is easier to see the spectral line, and to compare the emission in different channels, when one doesn't have to filter out (mentally or algorithmically) the emission common to all channels. Subtracting the continuum minimizes, and in some cases avoids entirely, the effort involved in deconvolving

introducing the delay, which can then easily be done correctly only for a single frequency. See Lecture 2.

the line signal, since it removes the need to deconvolve the same continuum emission from every channel. This is a particular benefit because deconvolution is inherently non-linear, and may give rather different results for different channels simply because the (u, v) coverage and the noise are different. For similar reasons, a better continuum image will result from combining the data from all the line-free channels before deconvolution.

This section discusses the four main methods currently used to subtract the continuum from the line data. Following the papers in this field and AIPS/Miriad usage, I call these IMAVG, UVSUB, UVLIN, and IMLIN. *This discussion however is a general one, not intended to reflect the details of any specific implementation.* For parallel discussions and comparisons of these methods, plus some practical hints, see the *AIPS Cookbook*, the *Miriad User Guide*, and especially Neil Killeen's *Analysis of ATCA Data with AIPS*.

6.1. IMAVG

The simplest method of removing the continuum is to make the spectral line data cube, and subtract from it a dirty image made by combining the data from all the line-free channels. This method has two major shortcomings.

(1) It ignores the change of the (u, v) coverage (and hence of the beam) with frequency. Subtracting the same continuum image from each channel will at best remove the continuum perfectly at one frequency, with errors due to chromatic aberration increasing away from that frequency (cf. Figure 12–4). This does not matter if the total bandwidth is small compared to the observing frequency ($\Delta\nu \ll \nu_0$), so the beam does not change much over that bandwidth; or if the continuum emission is relatively weak, so that the changing sidelobes are unimportant.

(2) It does not take into account changes in the apparent source structure across the observed frequency range. The most obvious source of such changing structure is a non-zero spectral index, but over wide enough bands the (u, v) coverage may also change enough to sample different aspects of the source structure.

6.2. UVSUB

Both of the disadvantages of IMAVG may be alleviated by deconvolving the data from the line-free channels, and subtracting the Fourier transform of the resulting model of the continuum emission directly from the spectral line visibilities (see Ekers & van Gorkom 1984; van Gorkom & Ekers 1989). This automatically accounts for chromatic aberration (Figure 12–4), incidentally allowing for channel-dependent flagging as well. If the model is derived from data which span the observed frequency range, the different sampling of the true source structure will be taken care of as well. Such a multi-frequency deconvolution can also in principle allow for intrinsic changes of source structure with frequency, removing the effect of spectral indices and the like (see Lecture 21). Finally, this is the only method which can, when used carefully, correctly remove continuum emission spread over large areas on the sky, even outside the primary beam. This makes it the method of choice for the removal of simple (especially point) sources which are distant from the field center.

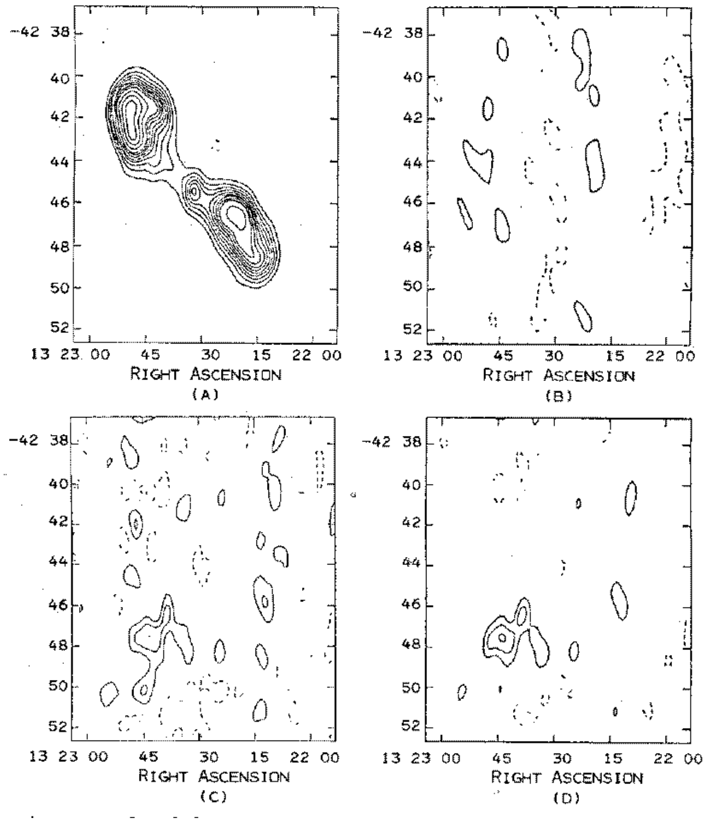


Figure 12–4. An example of the importance of proper continuum subtraction, and the effects of chromatic aberration. **(a)** shows the continuum source, the radio jet in the elliptical galaxy Centaurus A. The contours range from 1 to 18 Jy/beam. **(b)** shows the line-minus-continuum image at 260 km/s, where the dirty mean continuum image has been subtracted from the dirty channel image (the IMAVG method described in the text). The contour interval is 260 mJy/beam. **(c)** is the same as (b), but using the UVSUB method (i.e., subtracting the Fourier transform of the CLEAN model for the continuum from the channel data in the (u, v) plane). The contours are 26 mJy/beam — a factor 10 lower than in (b). **(d)** is (c) after deconvolution (with CLEAN); the H I is real! The total flux density of the continuum source is 200 Jy, which is why the residuals in (b) are so high that the line remains undetected. *Taken from van Gorkom & Ekers (1989).*

Unfortunately this method too has its disadvantages. Both the deconvolution to obtain the model, and the Fourier transforms (one for each channel) to subtract it, are very expensive computationally. The deconvolution step also requires some, and perhaps considerable, guidance from the observer. And since deconvolution is a non-linear process, and neither the (u, v) coverage nor the data are perfect, the derived model may not fully or correctly represent all the continuum structure. Any errors in the model will show up as systematic errors in the corrected spectral line cube.

6.3. UVLIN

UVLIN[6] and IMLIN are the most recent approaches to continuum subtraction (see van Langevelde & Cotton 1990; Cornwell, Uson, & Haddad 1992; Sault 1994, 1995), and have basically solved the problem for many simple experiments. Given their relatively recent appearance and widespread utility I discuss these methods at some length.

In the UVLIN scheme, a low (usually first) order polynomial is fit to the line-free part of each visibility spectrum. This polynomial is then subtracted from the entire spectrum, and possibly also saved as an estimate of the continuum visibilities. It is better to fit the real and imaginary parts of the data separately, since this keeps the whole process linear; fitting to the amplitude and phase also produces an amplitude bias at low signal-to-noise ratios.

The approximation that the continuum visibilities are linear (or a low-order polynomial) in frequency corresponds to restrictions in the total bandwidth observed, and in the distance of the continuum emission from the phase center. In one dimension, consider a point source offset from the phase center (since the process is linear, this suffices as well for more complicated distributions). The corresponding visibilities have constant amplitude and a linear phase gradient, and the real and imaginary parts vary sinusoidally with the visibility spacing:

$$V = \cos\frac{2\pi\nu b\, l_0}{c} + i\sin\frac{2\pi\nu b\, l_0}{c} \tag{12–4}$$

where b is the baseline length and l_0 is the offset of the point source from the origin (phase center). Figure 12–5 shows these visibilities for an H I observation covering a 6.25 MHz bandwidth centered on 1420 MHz, for two baselines corresponding to the range of those available in the VLA B configuration, and offsets of 15 arcseconds (four synthesized beams) and 15 arcminutes (the half-power point of the primary beam). For the source nearer to the phase center the linear approximation is excellent, but for the more distant source the approximation breaks down as the sinusoid becomes more evident. Using a higher order polynomial fit, as suggested by Sault (1994), extends the region over which the polynomial approximation is valid, hence enlarging the field-of-view over which UVLIN works.

UVLIN offers many advantages over more traditional methods of continuum subtraction.

- **Speed:** UVLIN is very fast, involving only a polynomial fit to each visibility spectrum.
- **Ease of use:** apart from deciding which channels are free of line emission, no user intervention is required.
- **Robustness:** As a linear process UVLIN is robust to most common systematic errors, such as antenna gain and some bandpass errors. Since each

[6] This method is implemented in AIPS in several different tasks, among them `UVLIN`, `UVBAS` (deprecated), `UVLSF`, and `UVMLN`. The reader is directed to the relevant `EXPLAIN` files for details and comparisons.

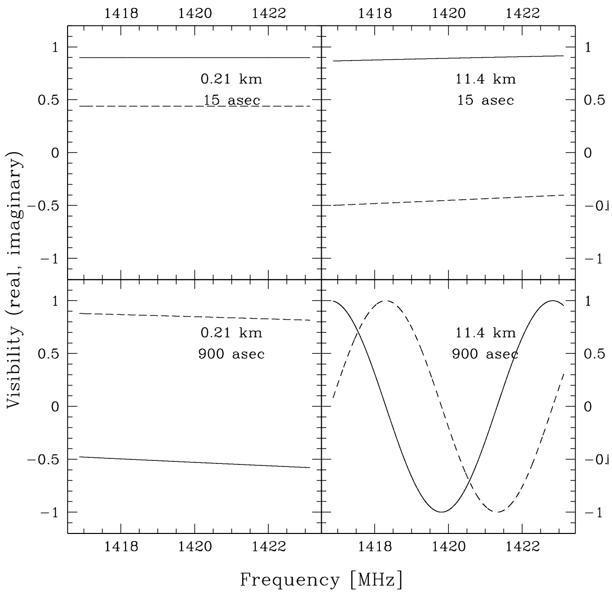


Figure 12–5. Visibilities as a function of frequency, for baselines corresponding to the range available in the VLA's B configuration. The real part of the visibility is shown as a solid line; the imaginary is dashed. The top panels illustrate the response to a point source of unit amplitude 15 arcseconds (about 4 synthesized beams) away from the phase center; the bottom panels are the corresponding plots for a unit amplitude point source at the half-power point of the VLA's primary beam (15 arcminutes from the phase center). The frequency range (6.25 MHz) is fairly typical for extragalactic H I observations. A linear fit would be a good approximation except for the most distant source on the longer baselines; this failure corresponds to bandwidth smearing (§5.) and the field-of-view limitation on UVLIN.

baseline is treated independently, any time variability will be removed automatically, whether it is intrinsic or due to instrumental effects such as primary beam rotation. Further, any residual gain errors impose a dynamic range limit on the line rather than on the continuum, an important point because the line is usually much weaker.

- **Spectral index:** UVLIN automatically corrects for the spectral slope of the continuum across the band.

- **Automatic flagging:** UVLIN naturally lends itself to automatic flagging based on the residuals in the line-free channels. The presumption is that a baseline which shows residuals much higher than expected (based on the weight and the thermal noise) in one or more line-free channels, is probably corrupted in channels with line emission as well. Hence a baseline should be flagged if its residual flux is above some cutoff level in any channel used in the polynomial fitting[7]. Since this flagging is done blindly on a very large number of baselines, it is best to be conservative, throwing out

[7] There are more sophisticated algorithms for finding bad data, e.g., checking the distribution of the residuals rather than simply the outliers, using the residuals on 'clean' baselines as an

only those baselines with residuals 7 to 8 times the expected rms noise level. This approach works so well in practice that it can seem almost magical, and often the data set needs no further editing at all. Because of this capability it has become common to use UVLIN even for continuum data sets, simply to avoid painstaking flagging by hand (cf. the AIPS task `FLGIT`).

- **Produces a good continuum data set:** The polynomial fit, together with the automatic flagging, provides a good continuum data set for subsequent imaging, if desired.

- **Analytic error estimates:** The magnitude and form of the residuals may be estimated analytically, giving some confidence limits on the resulting images (see Sault (1994) and Miriad's `conterr`). Basically the errors in the continuum subtraction will be greatest where the continuum is large, and where the point spread function (PSF) is also large.[8] More precisely, the worst errors for an nth order fit will occur where the $(n+1)$th derivative of the PSF is large (cf. Sault 1994).

Although UVLIN works beautifully for a great variety of experiments, it does have a few restrictions.

- **The channels used in the fitting must be entirely free of line emission.** The Fourier transform relationship spreads emission confined to a small spatial region over a large area in the (u, v) plane, so a channel with line emission *anywhere* cannot be used in forming the continuum. This is a major restriction in cases where the line is very broad and there are few or no line-free channels.

- **UVLIN assumes that the (u, v) coverage does not change much with frequency.** By default most implementations assume that the (u, v) coverage is identical for all channels, though this is not strictly necessary. What *is* required is that all baselines have at least some line-free channels to be fit.

- **UVLIN works well only over a restricted field-of-view.** The linear approximation fails for continuum sources too far from the phase center (see Figure 12–5). The restriction is identical to that for bandwidth smearing:

$$\theta \ll \frac{\nu_0}{\Delta\nu_{\rm tot}}\theta_S \qquad (12\text{–}5)$$

empirical check on that distribution, etc. While these are used for temporal flagging at some single-dish telescopes, notably those observing in the far infrared, little work has yet been done on applying them to interferometric data.

[8] There is some confusion in Sault (1994) as to the meaning of Cornwell, Uson, & Haddad (1992)'s equation 34. As Sault points out, the residual continuum errors must fall off with distance from a continuum source as the point spread function does. However, this does not invalidate the earlier paper's argument that the error *measured in units of the local rms sidelobes of the PSF* does not vary with distance from the point source. This contrasts with the case of IMLIN (q.v.), where the errors, measured in units of the PSF's sidelobes, grow quadratically with distance from the continuum source.

where ν_0 is the observing frequency, $\Delta\nu_{\rm tot}$ is the total bandwidth observed, and θ_S is the size of the synthesized beam (Cornwell, Uson, & Haddad 1992, eq'n 29). The residual continuum depends on the continuum flux density S_ν and on η, with

$$\eta = \frac{1}{2}\frac{\Delta\nu_{\rm tot}}{\nu_0}\frac{\sqrt{l_0^2+m_0^2}}{\theta} \tag{12–6}$$

where θ is the size of the synthesized beam for point sources, or the size of typical image features for extended ones, and (l_0, m_0) is the position of the source. Sault (1994) gives the approximate peak continuum residual for linear fits as

$$S_\nu \frac{\pi^2}{9}\eta^2 \tag{12–7}$$

Note that the residual continuum error is proportional to the distance of the point source from the phase center (Sault 1994). The field-of-view may be extended in four ways:

(1) By fitting higher-order polynomials: for a polynomial of order n, the residual scales as $\eta^{\rm n+1}$ (Sault 1994). Unfortunately higher-order fits are more susceptible to noise and other errors, and one must take care that such higher-order fits do not introduce artificial features in the channels excluded from the fits. This is not too much of a problem if one has a large number of line-free channels on both sides of the line emission.

(2) By using UVSUB to remove the more distant sources before running UVLIN.

(3) By shifting the phase center of the data before doing the fit. Since the field-of-view for UVLIN is centered on the phase center, one should shift the phase center to the position of the strongest continuum source before doing the polynomial fit. Unfortunately one cannot simply shift to one source, subtract it, then shift to and subtract the next, etc., because the fit at the initial position will have removed some of the flux from sources at the other positions, and the remaining residuals of those sources can no longer be fit by a low-order polynomial.

(4) By fitting low-order polynomials to emission from two positions simultaneously. This avoids the introduction of higher-order terms by explicitly fitting for two separate sources at once. It is also more reliable than fitting with a higher-order polynomial, firstly because that approach simply will not work for very distant sources, and secondly because one is explicitly including extra knowledge in the algorithm (the fringe rate of the other continuum source), which hopefully allows one to fit fewer parameters. This approach has been advocated by Sault (1995) for removing solar interference, and is available so far only within Miriad's `uvlin`.

6.4. IMLIN

The IMLIN method can be thought of as an image-based parallel to UVLIN. One subtracts a low-order (usually linear) polynomial fit to the line-free portion of the spectrum measured at each (spatial) pixel in the spectral line image cube. This algorithm shares most of the characteristics of UVLIN, primarily because both are linear algorithms. Like UVLIN, IMLIN is fast, easy to use, and robust to gain errors and spectral index variations. Like UVLIN, it requires the fitted channels to be completely free of line emission (because the PSF spreads physically localized emission over a wide area of the image), and it has a similarly restricted field-of-view. There are however a few significant differences.

- **Different error properties:** As pointed out in Cornwell, Uson, & Haddad (1992), IMLIN's errors scale somewhat differently than those of UVLIN. For IMLIN, equation 12–7 still holds, but with

$$\eta = \frac{1}{2}\frac{\Delta\nu_{\rm tot}}{\nu_0}\frac{\sqrt{(l-l_0)^2+(m-m_0)^2}}{\theta} \tag{12–8}$$

  where (l, m) represent the position of the location of interest. This means that the errors for IMLIN scale with distance from the continuum source, rather than with distance from the phase center. This difference in error patterns is sometimes helpful in distinguishing real emission from artifacts from the continuum subtraction; *it also implies that IMLIN is better at removing point sources at large distances from the phase center.*

- **Cannot flag data:** Since IMLIN works in the image plane, it cannot be used to flag the (u, v) data directly. This is probably the main practical defect of the algorithm, and the most important reason observers tend preferentially to use UVLIN.

- **May be better for excellent (u, v) coverage:** Sault (1994) points out that IMLIN, working after the Fourier transform with all the continuum signal concentrated in a few pixels (usually) rather than spread across the (u, v) plane, may be more robust and less noisy when one has excellent (u, v) coverage. This is not entirely obvious and remains to be shown analytically.

Whether UVLIN or IMLIN is better depends on the observation. Probably it is wise to use both and compare the results.

6.5. Errors in Continuum Subtraction

While continuum subtraction is almost always necessary and generally produces much better images, one unfortunate side effect is the transfer of noise from the line-free to the line channels. If the original noise level in a single channel is σ_0, the noise level in a line channel after subtracting a continuum estimated from N channels will be roughly

$$\sigma \approx \sigma_0\sqrt{1+\frac{1}{N}} \tag{12–9}$$

(this assumes a zeroth-order [average] fit; see Sault 1994 for a more detailed discussion). Note that the noise in the channels from which the continuum was estimated will *drop* by a similar amount. The noise in the line channels increases because a continuum estimate based on imperfect data is itself noisy. What is worse, this additional noise is systematic, since the same (or a closely related) continuum is subtracted from every channel. Further, for higher-order fits (linear and beyond) the errors in the continuum increase as one moves away in frequency from the channels which were used in its estimation. This increase becomes even more severe when one extrapolates beyond the range of line-free channels. The bottom line is that good continuum subtraction requires a large number of line-free channels spread roughly evenly on either side of any spectral features. How important continuum errors are for your experiment depends on the strength and location of the continuum. Unfortunately line emission often sits on top of or near continuum sources, and subtraction errors can readily create a false line or mask a real one (e.g., Figure 12–4).

Finally, note that bandpass errors are the most common causes of problems with continuum subtraction. While UVLIN and IMLIN are fairly robust to errors in the overall gain, and even to slopes in the bandpass, nothing can save a data set in which bumps and wiggles in the frequency response have not been calibrated out.

7. Flagging

Although spectral line observations avoid many systematic errors that affect continuum data, they are also vulnerable in ways that those data are not. Narrow channels do help minimize the effects of bandwidth smearing (see §5.) but this is a curse as well as a blessing, because it leaves spectral line observations vulnerable to the sidelobes of distant sources (e.g., the Sun) which would otherwise cancel out. Similarly, narrow RFI spikes may make spectral line channels unusable, while the frequency averaging inherent in continuum observations will minimize their effects. This is a particular problem because one can seldom choose at which frequency to observe a given spectral line; continuum observers can tune around bad regions of the spectrum, but line observers have no such luxury. For the same reason line observers will sometimes operate at frequencies the telescope designers did not envision, at or beyond the nominal band edges, particularly for highly redshifted lines. This can result in loss of sensitivity, complicated bandpass shapes, and simple failures, such as the LOs failing to tune or to lock and various filters cutting in.

These problems make careful flagging more important for spectral line than for continuum data. However, the sheer size of the data sets makes manual flagging more difficult — not only are there many channels to check (often a hundred or more), but the time spent on-source may be long, since one cannot increase the spectral sensitivity by using a wider bandwidth. For this reason, and to avoid having (u, v) coverage which changes significantly from channel to channel, a number of short-cuts are often employed:

- **Flag based on the pseudo-continuum:** In many bands most bad data are due to broad-band problems associated with individual telescopes or baselines, rather than single frequencies. Hence it is sensible to flag all channels based either on a single channel, or on a ‘pseudo-continuum’ (the VLA’s channel 0) formed by vector-averaging the central frequencies of the line data.

- **Check the bandpasses:** The bandpasses provide straightforward diagnostics of serious channel-based problems, which will show up as sharp dips or spikes. For this reason some observers compute bandpasses based on their line (as well as the usual calibrator) sources.

- **Automatic flagging:** As discussed above, algorithms like UVLIN often do a surprisingly good job of flagging data based on deviations from the expected spectral behavior. Similar automated flagging could be done in the time domain, though none of the standard reduction packages allow for this. The closure phase provides a useful diagnostic during bandpass calibration; for bright sources, one may also examine the decorrelation (ratio of vector- to scalar-averaged visibilities) over the band, which also checks the bandpass calibration. Eventually, as correlators produce more and more channels and RFI becomes more prevalent, RFI excision will have to be done on-line (see Fisher 1996 for some thoughts on this in the context of the GBT).

In the end one may be forced to examine data from each of the individual channels, but this is very much a last resort.

One point to bear in mind while flagging is that, for line observations, the (u, v) coverage for all of the channels should be kept as similar as possible. This is not necessary for continuum data taken in multiple channels, but it is very helpful when comparing the line emission at different frequencies, and even more helpful when analyzing such data. The danger is that changes in the (u, v) coverage produce changes in the point spread function, which may mimic or obscure real changes in the spectral line. Deconvolution should help, but as a nonlinear process may itself create substantial errors simply due to a changing (u, v) coverage.

8. Data Quality and the Spectral Dynamic Range

One measure of the ultimate quality of a spectral line cube is the *spectral dynamic range.* This is commonly used to mean at least three different things:

1. The ratio of the peak continuum signal to the root-mean-squared (rms) noise in a continuum-subtracted image. A high spectral dynamic range (SDR) in this sense corresponds to having a very flat frequency response (allowing the continuum to be accurately removed) and low thermal and systematic noise.

2. The ratio of the peak continuum signal to the accuracy with which one can measure a very deep absorption line. A high SDR in this sense corresponds to correctly measuring a very wide range in correlation coefficient.

3. The ratio of the peak of a monochromatic signal to its spectral sidelobes. A high SDR in this sense corresponds to very little cross-talk between frequency channels.

All three reflect the difficulty of an observation is, but spectral line observers almost always use the first definition, and that is what I will mean hereafter. Observations of faint lines superposed on strong continuum signals are difficult; the line-to-continuum ratio in a recombination line may be only a few per cent, demanding an SDR throughout the image of 1000:1 or more. Apart from the source strength and the thermal noise, which have so far limited millimeter interferometers to SDRs of a few hundred to one (L. Testi 1998, priv. comm; M. S. Yun 1998, priv. comm.), the main difficulty lies in determining the bandpass well enough to eliminate any leakage from the continuum into the line images. At the VLA the time-variable bandpass limits the SDR to perhaps 1000:1 without careful calibration; however, the best so far achieved with this instrument is more than 10^5 to 1 (van Gorkom *et al.* 1993), without having to remove baseline-dependent errors. Residual bandpass errors, and the correspondingly incorrect continuum subtraction, are particularly troublesome because they are often systematic with frequency, and can easily mimic real spectral lines.

9. Deconvolution

The deconvolution of spectral line images, while conceptually similar to that of continuum images, in practice presents rather different problems. (The problem of imaging continuum data taken in spectral line mode – multifrequency synthesis – is dealt with in Lecture 21.) These stem from the desire to compare line emission in different channels, while expecting that emission to look different, in strength and distribution, in those channels. Figure 12–6 illustrates some of the difficulties for a typical H I observation. This particular galaxy is the face-on spiral NGC 1058, as observed by the VLA in C+CS+D configurations. The emission is intense and widespread in some of the central channels, where we sample the flat part of the rotation curve, but becomes both fainter and more concentrated in the edge channels. Furthermore, although there are regions of high surface brightness, there are also large areas containing real but faint (in each pixel) emission. Since an interferometer responds very differently to structures at different size scales, the quality of these dirty images varies with frequency according to the mismatch between the observed spacings and the intrinsic structure. This is most obvious as the missing short spacing problem: in the channels with the most extended emission that emission sits on top of a negative 'bowl', while those with relatively compact structure show no such effect. This is particularly unfortunate because the error is systematic; the flux densities will be most severely underestimated near the line centers, where the emission is strongest and most widespread. These errors will show up in the analysis as, for instance, systematic overestimates of the velocity dispersion. Worse yet, the extent of this effect changes with position in the galaxy, since the emission at each location covers a different range of channels, and since the negative 'bowl' in each channel is not constant across the image (Figure 12–7).

Deconvolution for line data is important for three related but rather different reasons. First, deconvolution removes the sidelobes of bright sources, which can dominate the noise in an image. For continuum images this is generally its most important role, since the sidelobes of a few very bright sources can often obscure fainter structure in the dirty image. Observations of masers similarly bene-

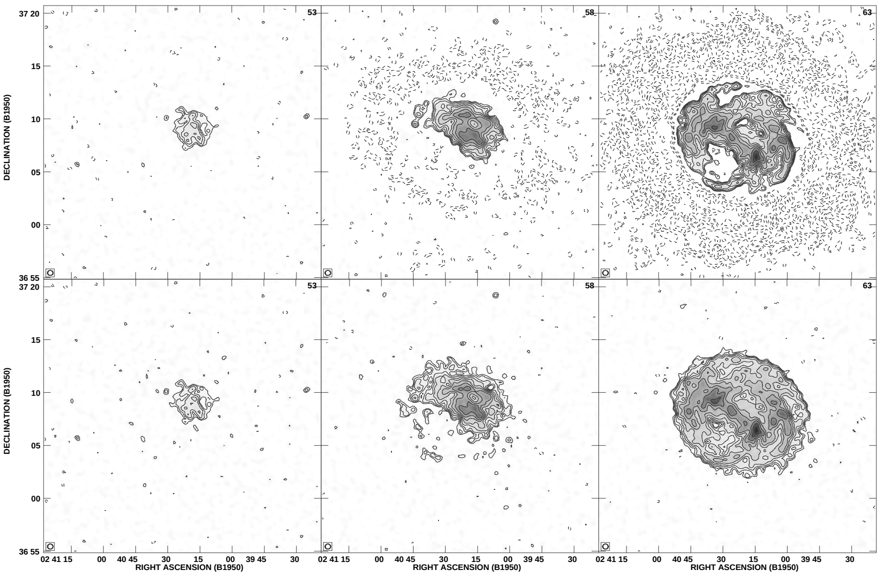


Figure 12–6. H I in the face-on spiral galaxy NGC 1058: a sampling of channel maps. The top row shows dirty maps, the bottom row CLEANed images. Note the large negative bowl in the dirty maps with the most extended emission. Contours are $\pm 2^{n/2}\sigma$, where σ is the rms noise (0.585 mJy/beam) and $n = 3, 4, \ldots$.

fit from the removal of the sidelobes of a few bright, simple sources. However, for the extended, low surface brightness emission characteristic of thermally-excited lines this function is almost negligible, because there are usually no small, dominant sources. Thus it has often been said that one need not deconvolve H I images, or at least not CLEAN very deeply, because the peak emission may only be two to three times the noise level, and the sidelobes of such weak emission are clearly negligible. This would be true for a few weak point sources, but the sidelobes of extended emission taken as a whole may be formidable even if the sidelobes of any single pixel are unimportant.

The second function of deconvolution then is to remove the integrated sidelobes of low-level but extended emission; most obviously, to remove the negative 'bowl' discussed above. This is closely related to the third function of spectral line deconvolution, to allow the measurement of flux densities, and particularly the total flux density. Since the integral of the dirty beam is zero (interferometers do not measure the zero spacing, and so are not directly sensitive to the total flux), one has to deconvolve an image — replace the dirty by the restoring beam, usually chosen to be nicely positive — in order to estimate the flux density within a given region (Figure 12–8). One can get away with ignoring this (to some extent) for small sources by integrating over the smallest possible region, since most of the positive area of the dirty beam sits near its central peak, but for large sources this is simply not possible. This brings up another subtle but crucial point, that the flux density scale of the restored model is different from that of the residuals, since the one is restored with the positive, Gaussian CLEAN beam, while the other is still (roughly speaking) characterized by the

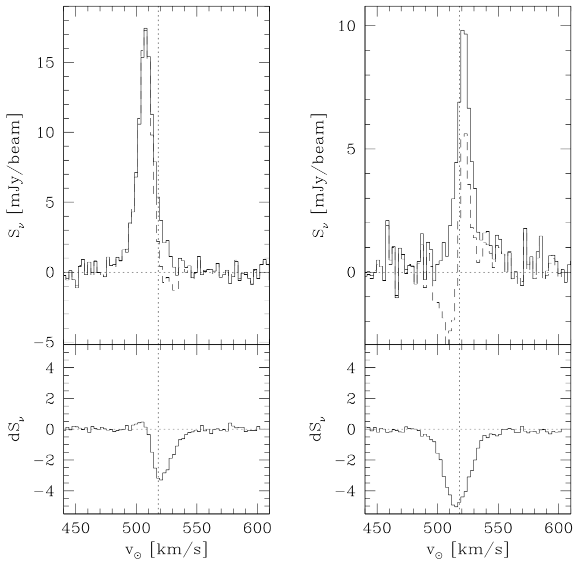


Figure 12–7. H I spectra through random positions in NGC 1058. The top panels compare the spectra taken from the dirty (dashed) and the CLEAN (solid) cubes; the bottom panels show the differences. The vertical dotted lines indicate the systemic velocity. The dirty maps systematically underestimate the flux, but do so by different amounts in the different channels, biasing estimates both of the bulk (rotational) velocities and of the line widths. The amount and sense of the bias depends critically on the spatial structure of the emission in each channel.

original, complicated dirty beam. These problems occur for continuum as well as line data, but spectral line observers seem to be more sensitive to them. In part this is because of differences in source structure and the desire to compare channel maps with radically different emission characteristics on the same footing. But there is also the practical consideration that deconvolving hundreds of channels takes a long time, and observers not unnaturally would prefer to do as few CLEAN/MEM iterations as they can get away with.

Unfortunately there are no good algorithms commonly available for handling faint, extended emission. CLEAN is terribly inefficient for such problems because it models the emission as a sum of point sources. Not only does it take forever (CLEANing to 1σ might easily take tens of thousands of iterations for a single channel[9]), CLEAN also changes the noise characteristics of the data, making the image artificially 'pointy'; furthermore the off-source noise should be taken only as a lower limit to the on-source errors (cf. Cornwell, Holdaway, & Uson 1993). Various attempts to modify or extend the CLEAN algorithm to cope with smooth, extended structure (notably smoothness-stabilized CLEAN, Cornwell 1983, and multi-resolution CLEAN, Wakker & Schwarz 1988 – see also Braun 1995) have not been very successful in general, although they do help in some specific cases. The usual lore is that MEM algorithms should do

[9] The Steer, Dewdney, & Ito (1984) version of CLEAN speeds these deep CLEANs considerably.

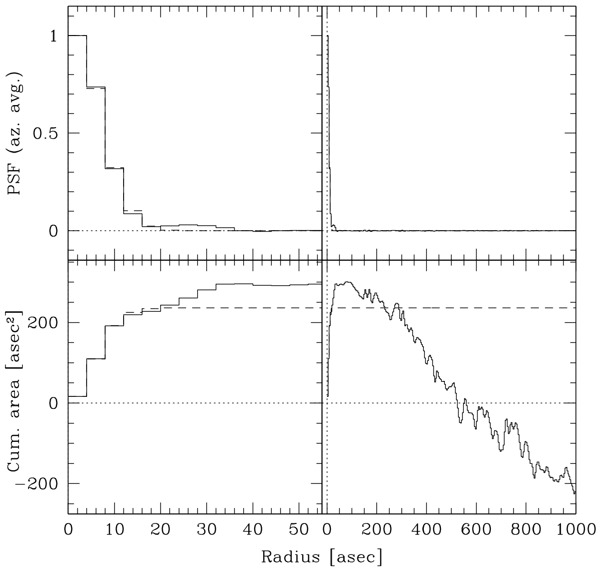


Figure 12–8. Comparison of dirty (solid lines) and CLEAN (dashed lines) beams, for a three-configuration (C+CS+D) H I study of NGC 1058 (Rupen & Petric, *in prep.*). The top panels show the azimuthal average of the beam (point spread function) *vs.* radius, while the bottom panels display the cumulative area. The dirty and the CLEAN beams match very well at small radii, but although the sidelobes are small (less than 2% beyond 60 arcsec), the cumulative area diverges rapidly beyond about 20 arcsec. The flux density is measured in Jy/beam, and this mismatch in beam areas creates significant problems when calculating total flux densities, or when adding residuals to a CLEAN or MEM model.

a much better job on such smooth, extended structure. Unfortunately, this is not the case (Rupen 1997). The usual implementations, such as the AIPS task VTESS, get much of their power from enforcing positivity; this leads to a significant bias in cases where the emission in each pixel is only of order the noise level. "Maximum emptiness" algorithms are also available (e.g., the AIPS task UTESS), which do not rely on positivity, but these seem to be very sensitive to boxing, to be no faster than than CLEAN, and to give virtually identical results to CLEAN for real data (Rupen 1997).

How deconvolution should be done in practice depends very much on the scientific problem one wishes to address, and remains in any case something of an art — that is, the pundits disagree considerably in what they recommend. On a couple major points however there is fair agreement.

First, it is wise to use a similar, and probably the identical, restoring beam for every channel. The dirty beam changes from channel to channel, both because the effective size of the interferometer scales with frequency, and because of any channel-dependent flagging (e.g., due to RFI at some particular frequency). Choosing a single restoring beam however makes the subsequent analysis much easier.

> *Second,* carefully restricting the area of the deconvolved model (boxing) can help considerably, since deconvolution algorithms get much of their power from the model's finite support. Unfortunately this means boxing every channel individually.

There is considerable debate on one of the most important points, how deeply to CLEAN (or equivalently, for how many iterations to run MEM). In part this is because different science demands different strategies. If the main interest is to see where the emission is and get a rough velocity field, deconvolution may be a waste of time — even the dirty images may be sufficient. On the other hand, if the goal is to measure the precise shape of a line profile or to compute detailed line ratios, one may wish to CLEAN them very deeply, if only to see what effect this has on the results. The main 'rules of thumb' which float around the community — stop deconvolving when the sidelobes of the residuals lie below the thermal noise; keep iterating until the flux converges; CLEAN all your channel maps for the same number of iterations, or to a constant flux level — are all only applicable in certain restricted circumstances, and in fact are mutually contradictory. For H I data, I would tend to recommend CLEANing all channels to a constant flux level, some fraction of the rms noise; this is based on the desire to recover the zero-spacing flux accurately for extended emission, and from comparisons of moderately to 'infinitely' deep CLEANs, which suggest that such fairly deep CLEANs do help in modeling the true source structure (see Rupen 1997 for a detailed discussion of some of these tests).

The technique developed by Jörsäter & van Moorsel (1995) to recover the total flux density accurately from fairly light CLEANs illustrates some of the complexities involved in deconvolution. They point out that the main effect of deconvolving weak emission is to replace the central part of the dirty beam by a Gaussian beam, which for simple beams and emission at about the noise level basically corresponds to multiplying by a scaling factor. So one can estimate the total flux density S_ν as the sum of the CLEANed flux $C_{\nu,i}$ and a scaling factor α times the residual flux $R_{\nu,i}$:

$$S_\nu = C_{\nu,i} + \alpha R_{\nu,i} \tag{12–10}$$

where i gives the number of CLEAN components used. Since S_ν and α should not depend on how deeply one CLEANs, they can both be determined by CLEANing the same image twice, with a different number of iterations, and recording $C_{\nu,i}$ and $R_{\nu,i}$.[10] This procedure accurately recovers the total flux density for their observations while requiring only a few hundred CLEAN components.

Unfortunately this procedure is not a panacea. Although it can help in determining the total flux density (for well-behaved dirty beams), it cannot remove the sidelobes of the dirty beam, and the negative bowl (amongst other artifacts) remains. Since the CLEAN and the residual flux densities depend on the distribution of the flux and on the boxing used in the deconvolution, α may vary from channel to channel and from source to source (see Rupen 1997 for an

[10] See the EXPLAIN file for the AIPS task `IMAGR` for W. Cotton's alternative approach to determining α.

example). Finally, although multiplying the residuals by a constant may lead to a more accurate flux estimate, it will also change the apparent noise level, and that noise level will change with the depth of the CLEAN. Clearly this approach is very useful for certain problems; equally clearly it must be used with some care.

A more general approach is to try to minimize the problem before one begins, by making the dirty beam as similar as possible to the desired restoring beam. This requires clever manipulation of the weighting of the (u, v) data through tapers, robustness, super-uniform weighting, etc., which is not always easy in current software packages. But the potential benefits are great, in minimizing the systematic problems inherent in non-linear deconvolutions, in keeping the residuals and the CLEAN/MEM model on a common flux scale, and in keeping the number of CLEAN/MEM iterations needed to a minimum.

10. Looking at the Images

The first stage in data analysis should almost always be to look at the data: to see where the emission is, to get a feeling for noise, and to try to learn what you can from the cubes without the prejudice of modeling or statistics. The final stage will often be very similar, a quest for the perfect plot or display to share with others the insights that may have taken months for the observer to develop and explore. This can be difficult even for simple images; when one is trying to understand or communicate three-dimensional data, the problem is compounded. Fortunately a host of tools have been developed over the years to illuminate some of the most important aspects of spectral line cubes, and a number of these have been found useful enough, in a range of contexts, that they have become fairly standard representations.

Unfortunately in a book like this one cannot readily illustrate the most powerful (and fun!) three-dimensional representations of data cubes. Probably the most familiar of these is a 'movie' flipping through each of the channels in turn, a simple aid which is enormously helpful in spotting relations between channels, and in figuring out whether emission features marginal in one channel are actually there. Recently the visualization library Karma (e.g., Gooch 1996)[11] and commercial packages like IDL have made more sophisticated displays available; these really need to be used to be appreciated, but the possibilities include volume rendering of cubes from arbitrary viewing angles (Figure 12–9) and simultaneous 'movies' that present slices in position as well as in velocity.

Journals thus far have not embraced three-dimensional or interactive displays, so observers employ a wide variety of one- or two-dimensional substitutes. By far the most common approach to conveying all the information in a full cube is to show contour plots of the individual channels using adjacent plots for adjacent velocities. These are particularly useful when the Doppler shift determines the observed line frequency, so each channel corresponds to a 'slice' in velocity. Thus a galaxy (rotating disk) with a flat rotation curve looks something like a pair of scissors opening and closing (Figure 12–11), because the surfaces of

[11] Karma is available via public ftp at `http://www.atnf.csiro.au/karma`.

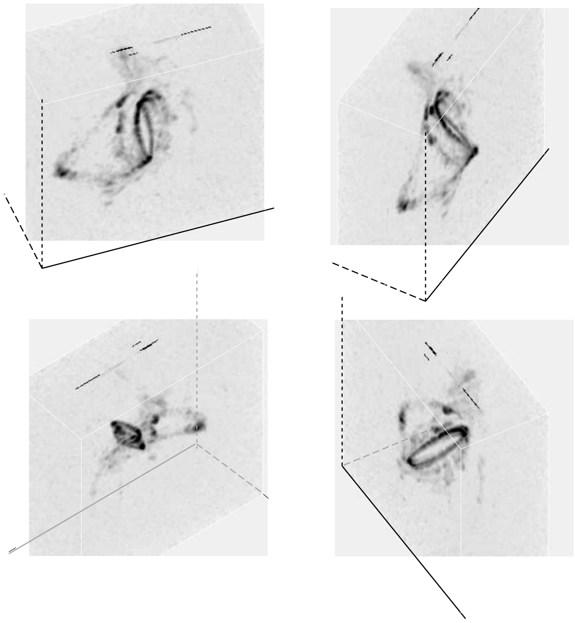


Figure 12–9. Volume rendering of a VLA H I cube of the M81 group (courtesy M.S. Yun; see Yun *et al.* 1994), made using the program `xray` in Karma. The wireframe shows the (α, δ, v) axes. The dominant, circular feature in the center is M81 itself. Note the peculiar almost helical structure seen in the upper right and lower left frames. The 'blob' of gas to the east, most obvious at the left of the top two images, is the companion NGC 3077; the somewhat dimmer grouping of gas north of M81 is the starburst galaxy M82, here under-represented because of absorption against its very strong continuum emission.

constant Doppler shift

$$v_c \sin i \cos \theta = \text{constant} \tag{12–11}$$

look roughly like parabolae; here i is the inclination, to first approximation constant for a given galaxy, v_c is the circular rotation speed, and θ is the position angle in the galaxy's disk with respect to the line-of-nodes (see Bosma 1978, 1981a, b; and van der Kruit and Allen 1978, for detailed discussions of the interpretation of galaxy velocity fields). Similarly, uniformly expanding thin shells show up as concentric 'donuts' of emission (Figure 12–13), filled spheres can be centrally peaked in every channel (Figure 12–14), and bipolar outflows appear as long tubes of emission pointing away from the central source (Figure 12–15). The important point here is that each channel selects a narrow range of velocities, which in turn is characteristic of a very localized spatial region. Thus one can use velocity information to reveal the full three-dimensional structure of the source. This is true so long as ordered motions dominate the line-of-sight kinematics.

The problem with these is that it is difficult to tell where a feature in one channel lies with respect to a feature in another — humans find it far easier

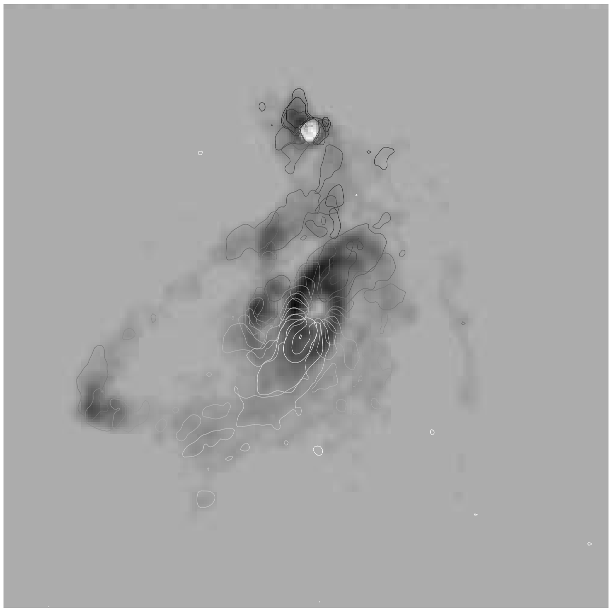


Figure 12–10. 'Renzo-gram' of the M81 H I cube (courtesy M.S. Yun) shown in Figure 12–9. The grayscale shows the total intensity (moment 0), with the white blotch to the north due to absorption against M82. Each contour corresponds to the same brightness level in a different channel, shaded between white and black depending on velocity; only every 6th channel is shown. The strong rotation of M81 is obvious, as is the kinematic connection between it and the H I streamers, and the latter's narrow velocity width. This sort of display would obviously be much better in color, e.g., red-to-blue contours on top of a grayscale intensity image. This particular plot was made with Karma's `krenzo`; Gipsy has a similar capability.

to extract information from a single, self-contained image, than from a series of adjacent ones. Hence the popularity of various forms of collapsing the three-dimensional line cube into two- or even one-dimensional displays. Moment maps are calculated as intensity-weighted sums across velocity, yielding images of intensity, velocity, and occasionally velocity dispersion as a function of position (Figures 12–12, 12–18). Slicing or summing along one spatial dimension produces *lv*-diagrams useful for showing rotation curves (Figure 12–12), finding expanding shells (which appear as double-valued velocities at a given position), and the like. One very useful plot is the so-called 'Renzo-gram'[12], in which contour plots of all the channels, using a single, constant contour level, are superposed in a single image (Figure 12–10). This is an excellent approach to displaying full velocity information when moment maps do not suffice, e.g., when there are multiple velocity components along a single line-of-sight.

[12] Named after Renzo Sancisi, one of the great interpreters of interferometric H I data.

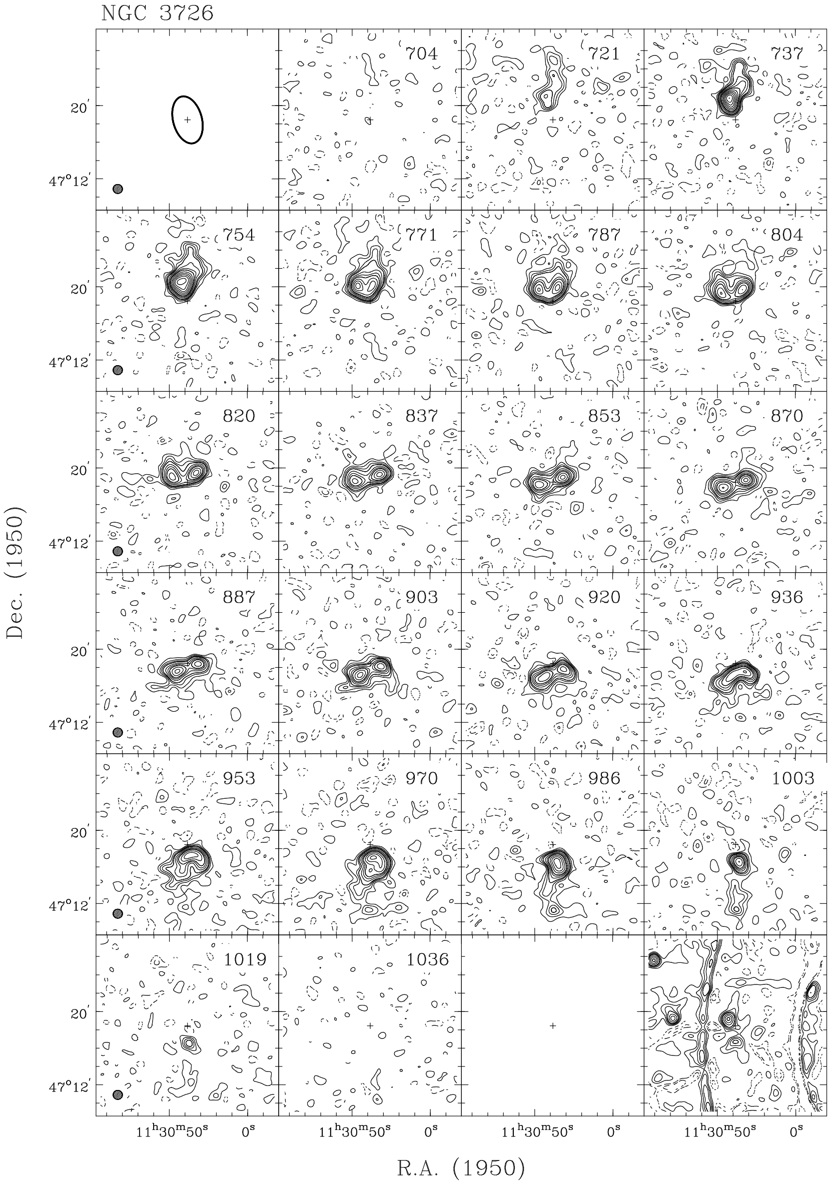


Figure 12–11. Channel maps of H I in the spiral galaxy NGC 3726 (Verheijen 1997), showing the H I as a function of velocity (marked in km/sec in the upper right-hand corners of the plots). The contours are −3, −1.5 (dashed), 1.5, 3. 4.5, 6, 9, 12, 15, ... times σ, where σ is the root-mean-squared (rms) noise. The synthesized beam (full-width at half-maximum) is indicated in the lower left corners of the left-hand panels. The upper left panel represents the optical galaxy, while the lower right panel shows the (dirty) radio continuum image. The emission moves from north to south as the velocity increases, becoming broader (east-west) in the central channels. This behavior is characteristic of a rotating disk. The asymmetry of the faint emission at large radii (e.g., 721 and 986 km/sec) suggests a warp in the gaseous disk beyond about 3 arcmin. The almost vertical 'rings' in the continuum image are the sidelobes of strong sources outside the displayed area; these rings are characteristic of east-west arrays such as the WSRT, where these data were taken.

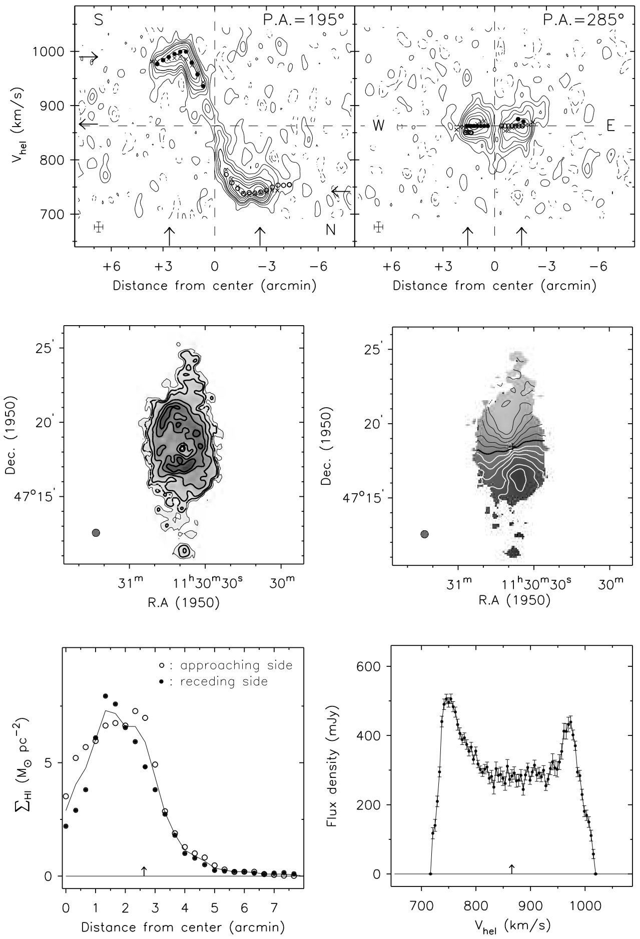


Figure 12–12. H I in NGC 3726 (Verheijen 1997). These are various alternative displays of the data shown in Figure 12–11. The **upper panel** shows two slices through the cube, along the major and minor kinematic axes — that is, the H I emission as a function of velocity and radius, along a line through the dynamical center of the galaxy at the indicated position angle. The major axis slice (left), often called an *lv*-diagram, is dominated by the galactic rotation, which produces the outer envelope to the observed emission. Purely circular orbits, uniformly populated with gas, would give an *lv*-plot which is point-symmetric about the dynamical center (marked by the dashed cross). The slight asymmetry in the emission within about an arcminute of the center may indicate a central bar (non-circular orbits). The corresponding minor-axis slice (right) would for circular rotation also be point-symmetric, with the ridge-line aligned exactly parallel to the x-axis if the gas at all radii were moving in the same plane. The slight deviations at large radii may correspond to the warp mentioned in Figure 12–11. The **middle panels** show the integrated surface brightness (left) and velocity field (right). White contours represent approaching velocities, and dark contours receding ones. The 'squashed spider' shape is characteristic of a rotating disk; the fact that some of the contours are closed indicates that the apparent velocity falls somewhat in the outer parts of this galaxy; while the distortion of the outer contours is the sign either of spiral arms or (more likely) a warp. Finally, the **lower panels** show the radial distribution of the H I (left), derived from the image of the integrated surface brightness, and the integrated H I spectrum of the galaxy, obtained by summing up all the emission in each channel map. The double-horn profile corresponds to the 'piling up' of emission in the edge channels (see the *lv*-plot in the upper left panel), and is a sign of the galaxy's overall rotation.

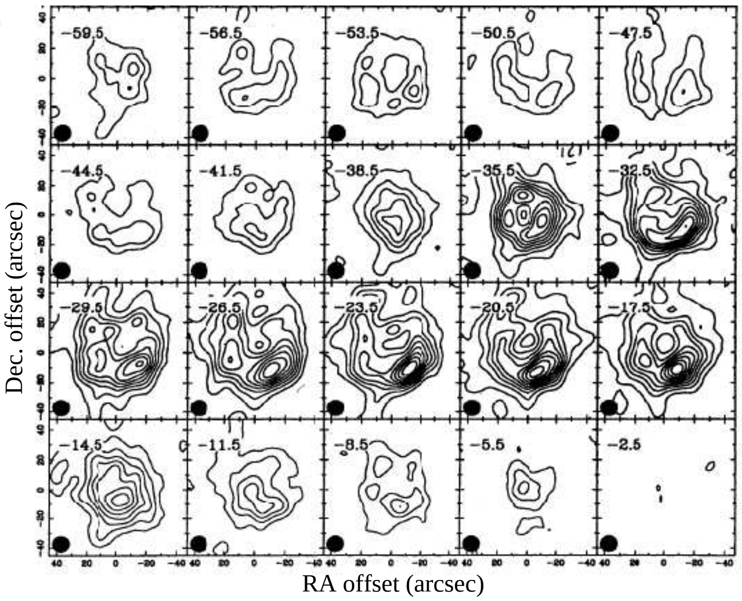


Figure 12–13. Contour images of CN (N=1-0) emission in the circumstellar material around IRC+10216, from Dayal and Bieging (1995). The change in apparent size with velocity, and the fact that most channels show central depressions, strongly suggest that this is an expanding shell.

11. Common Analysis Problems

After calibrating, imaging, and examining the resulting spectral line cubes, the real work begins: figuring out what all these lovely images are telling you about the universe. This analysis is limited only by the astronomer's imagination and (hopefully) good sense, but there are a few problems which are common enough to warrant discussion even in an introductory Lecture.

11.1. Distribution, Velocity Field, and Line Width

Probably the most common reason to observe a spectral line is to find out how much emission there is, and how it is distributed. For simple optically-thin lines, like H I, this is related directly to the column density along the line-of-sight; in other cases the lines are used simply as tracers of the special conditions needed to excite those transitions (e.g., the use of HCN to trace dense cloud cores, or certain maser lines to trace shock interactions). Spectral lines are also commonly used as kinematic probes, most famously for determining H I rotation curves of galaxies, but also in measuring the size-line width relation in molecular clouds, or simply as another handle on the three-dimensional structure of the gas. The applications are diverse, but all require estimates of the integrated line intensity, the velocity of each component, and possibly the line width, as functions of position.

The estimates in most common use are the moments of the line profile:

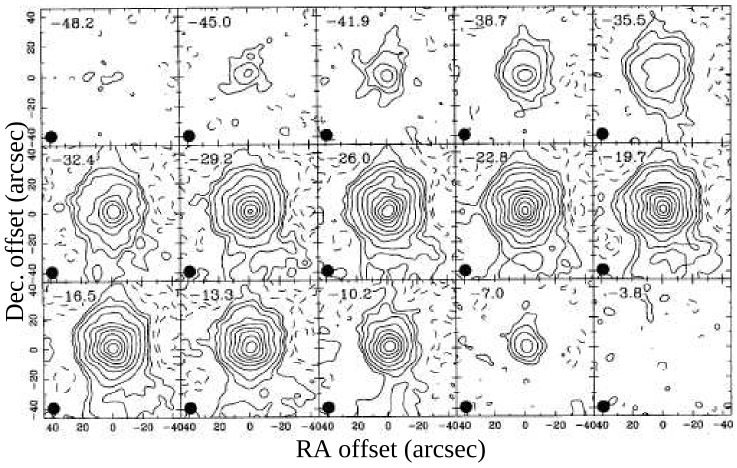


Figure 12–14. Contour images of HCN (J=1-0) emission in the circumstellar material around IRC+10216, from Dayal and Bieging (1995). Again the source changes size with velocity, but in this case most channels appear centrally peaked; this is characteristic of an expanding, filled sphere. The difference between this and the CN images shown in Figure 12–13 provides useful information on molecular chemistry in stellar outflows.

$$I_{\rm tot}(\alpha,\delta) = \Delta v \sum_{i=1}^{N_{\rm chan}} S_\nu(\alpha,\delta,\nu_i) \tag{12–12}$$

$$\overline{v}(\alpha,\delta) = \frac{\displaystyle\sum_{i=1}^{N_{\rm chan}} v_i\, S_\nu(\alpha,\delta,\nu_i)}{\displaystyle\sum_{i=1}^{N_{\rm chan}} S_\nu(\alpha,\delta,\nu_i)} \tag{12–13}$$

$$\begin{aligned} \sigma_v(\alpha,\delta) &\equiv \sqrt{\langle (v_i - \overline{v}(\alpha,\delta))^2 \rangle} \\ &= \sqrt{\frac{\displaystyle\sum_{i=1}^{N_{\rm chan}} (v_i - \overline{v}(\alpha,\delta))^2\, S_\nu(\alpha,\delta,\nu_i)}{\displaystyle\sum_{i=1}^{N_{\rm chan}} S_\nu(\alpha,\delta,\nu_i)}} \end{aligned} \tag{12–14}$$

Here $S_\nu(\alpha,\delta,\nu_i)$ is the observed flux density at the position (α,δ) in channel i (corresponding to frequency ν_i and velocity v_i); Δv is the velocity width of

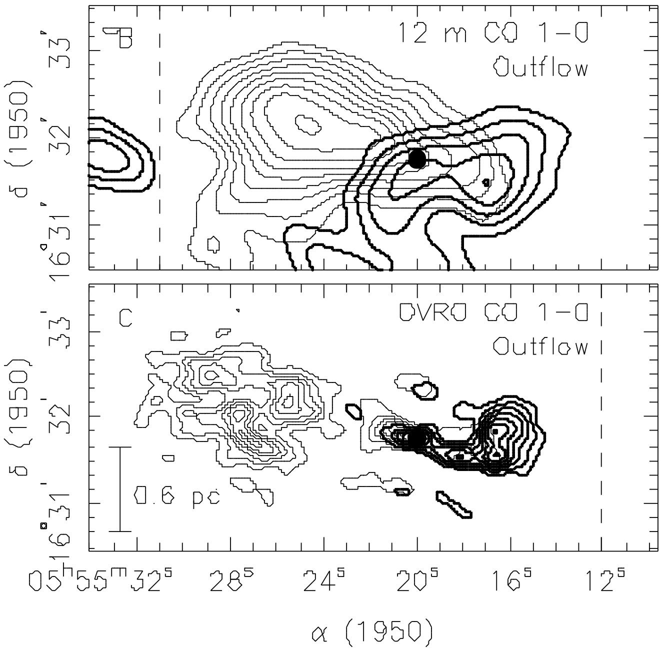


Figure 12–15. Superposed contour plots of blueshifted and redshifted molecular (CO) emission in G192.16, a bipolar outflow from a massive young star (Shepherd *et al.* 1998), taken with the NRAO 12m (single-dish, top) and OVRO (interferometer, bottom). The collimation and the symmetry of these edge (highest velocity) channels about a central core both suggest an outflow (jet).

the channels, assumed constant[13]. We thus collapse the three-dimensional cube $S_\nu(\alpha, \delta, \nu_i)$ into two-dimensional images of the information we are actually interested in, the integrated emission $I_{\rm tot}(\alpha, \delta)$, the intensity-weighted mean velocity $\overline{v}(\alpha, \delta)$, and the intensity-weighted velocity dispersion $\sigma_v(\alpha, \delta)$. These moments offer the advantages of being conceptually simple, intuitive, and well-defined. In practice however they can be difficult to use, for a variety of reasons (see Figure 12–16).

- **Sensitivity to clipping**
  The simplest approach to calculating the moments is simply to apply the definitions above to the entire cube. The problem is that at any given pixel the line signal is generally present in only a few channels; adding all the channels together thus includes many channels which contribute noise, but no useful information. The resulting moment maps can be much noisier than the individual channels (Figure 12–17; compare Figure 12–6). The higher-order moments become progressively more vulnerable to outliers,

[13] This is not precisely correct in most cases, since the channels generally have a constant width in frequency, corresponding to a different velocity width across the observing band; for most current observations this is not terribly important.

since the noise is weighted up by the velocity or the velocity squared (cf. equation 12–14), and even for very strong lines the results can be effectively meaningless, as illustrated in the left-hand column of Figure 12–18.

It would be much better to use only channels with line emission in the sums. The middle columns of Figures 12–17 and 12–18 show the improvement gained by simply discarding any points below twice the rms noise level (σ) in the images. Two major problems however remain. First, since the clipping is at a constant level, positive noise spikes are allowed in, while negative ones are discarded; this introduces a positive bias in all the even-numbered moments – most importantly total intensity and velocity dispersion. One could choose a higher cutoff, but this would also discard weak but real signals; or one could clip by absolute value to avoid the bias, but this introduces pixels which are clearly just noise. The second difficulty lies with true emission which is (in a single pixel) weaker than the cutoff and should be included for any realistic estimate of the total flux.

These considerations led to the modified approach which is most commonly used today (Bosma 1981; van der Kruit and Shostak 1982). The major improvements are (1) to smooth the data before clipping, and (2) to allow some 'spillage' in frequency, i.e. to use even low-intensity pixels in calculating the moments if those same pixels are above the threshold in neighboring channels. Smoothing the data both gives a higher signal-to-noise ratio, allowing fainter and more extended emission into the analysis, and provides some extra pixels around the edge of the source where one might expect real emission which is too faint to detect directly. Similarly, extending the 'good' regions of the cube to include pixels with emission at neighboring velocities allows for channels which may have real but very low-level emission. An example is shown in the right-hand columns of Figures 12–17 and 12–18. Here the original cube was smoothed down by a factor of two both spatially and in velocity; the smoothed cube was clipped at 3σ, with two channels alongside any $\geq 3\sigma$ pixel also counted as 'good.' The moments were then calculated from the original cube, using only the pixels which corresponded to good pixels in the smoothed cube. The resulting moment maps are both more consistent (from pixel to pixel) and more reliable, although the velocity dispersions remain somewhat suspect.

In any individual case one might prefer the intuition of the astronomer to the objectivity of an automatic procedure, and it may be useful to select regions of true emission interactively. This can be very powerful but apart from the subjective choices made can take a very long time if there are many channels.

- **Moments are not independent**
  The higher-order moments depend on those of lower order, for instance the velocity dispersion requires an accurate measure of the mean velocity.

This implies that the higher order moments become progressively noisier and more sensitive to exactly which pixels are used in their calculation.

- **Difficult to interpret**
  More fundamentally, the moments may not be useful in characterizing an arbitrary line profile. If both absorption and emission are present, summing the two together is not very meaningful. Interpreting the mean velocity and dispersion of a profile with multiple peaks (e.g., overlapping galaxies, an expanding bubble, etc.; cf. Figure 12–16) is difficult at best; this is even truer when several transitions are present in a single observing band. Moments *cannot* be used blindly — one must be well aware of the profile shapes, before using moments to characterize them.

- **Biased towards regions of high intensity**
  The moments are by definition biased against regions of low intensity. This is sometimes sensible, but may not be appropriate for some experiments. For instance, it is not at all clear that one is less interested in low-emission regions when modeling the dynamics of a galaxy — here the gas is used basically as test particles, and one may be interested precisely in those regions where the gas is most tenuous.

- **Hard to quantify and complicated error estimates**
  As argued above, simply using all the pixels when calculating moments leads to much higher noise levels than is truly necessary; it is far preferable to use only those channels which have real emission. Since the line profile in general changes across the source, this means that different numbers of channels will enter into the moment analysis at different pixels (Figure 12–18), and hence the noise level may vary greatly across the resulting images. Furthermore, the error associated with the moments depends directly on the intrinsic line shape, and the higher-order moments depend on the velocity of the noise as well as its magnitude. All this makes it difficult to assign quantitative error bars to any but the zeroth moment (total intensity).

In summary, while moment analysis is probably a reasonable way to calculate the distribution of the total intensity, higher moments should be used very cautiously if at all, and with a healthy respect for the possible pitfalls.

If the use of moments is problematic, what are the alternatives? One obvious approach is to fit a specific profile shape, most commonly a Gaussian. This is a well-defined procedure which, if done properly, can give not only real error bars on the fitting parameters, but also an estimate of the overall quality of the fit, i.e. how likely it is that a given profile is indeed well described by the adopted functional form. For well-designed fitting routines clipping is not necessary, and one can allow fairly easily for multiply-peaked profiles. The difficulty of course is choosing a function which has some physical basis and which indeed adequately represents the data. One might assume for instance that a Gaussian profile should be a good approximation for a line in thermal equilibrium, but systematic motions – or simply multiple clouds – within each synthesized beam ensure that the situation is not so simple. Even if it were, most existing packages do not actually estimate error bars or the goodness-of-fit, and seldom handle

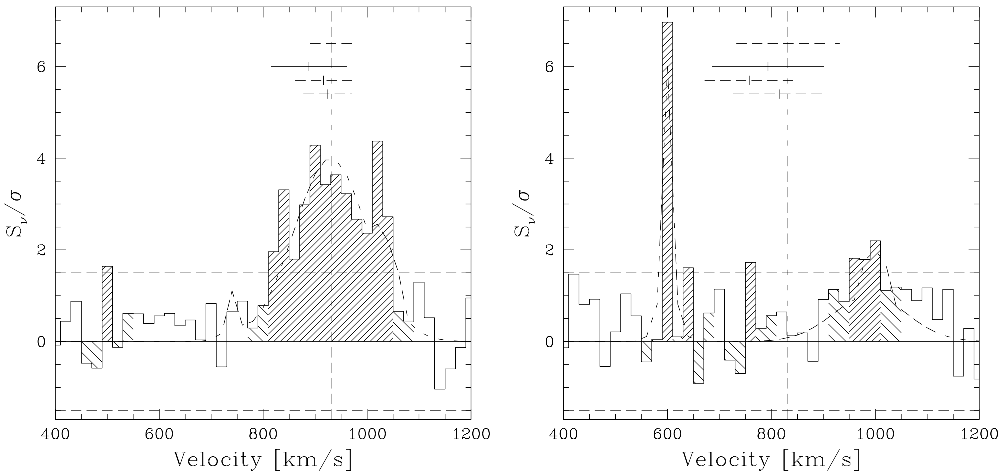


Figure 12–16. Sample spectra illustrating simple moment analysis. Dashed curves represent the true emission; the histogram shows the observations, binned and with noise added. The dashed vertical line shows the mean velocity (moment 1, $\overline{v}$) of the true emission, while the top-most dashed horizontal line indicates the velocity dispersion (moment 2, σ_v); the total intensity (moment 0, I) is 39.3/19.3 for the left- and right-hand spectra, respectively. Calculating moments from all channels with no clipping gives $I \approx 47/30$, and $\overline{v}$ and σ_v as indicated by the solid horizontal line and cross. Clipping at 1.5σ yields the dark-hashed region, giving $I \approx 38/16$ (more accurate but biased low) and the dotted horizontal bar/cross. Adding an additional 2 channels on either side of all channels above 1.5σ adds in the lightly-hashed channels giving $I \approx 40/20$, and the lower dashed horizontal line/bar. Clipping is clearly a good thing; smoothing and then clipping would do an even better job. Even so, the interpretation of higher-order moments for spectra with multiple peaks is problematic at best.

multiply-peaked spectra gracefully. Nevertheless careful modeling remains one of the best (as well as one of the most laborious) ways of extracting physical parameters from spectral line observations. Malhotra's (1994, 1995) derivation of scale-heights and velocity dispersions from Galactic H I and CO data is an instructive example, as are Olling's (1995, 1996a,b) parallel efforts in nearby galaxies.

Even without detailed modeling, moments are seldom the best estimators of characteristic velocities and line widths. More robust alternatives include the median rather than the mean for the velocity, and estimates based on the distribution of the flux (e.g., the difference between the first and third quartile, or the full-width at 10% of the peak) for the line width. Such estimators are commonly used to characterize single-dish profiles but have seldom been applied to interferometric data. Any statistic must be chosen with the observed profiles firmly in mind; the point is not to calculate some mathematical expression, but to characterize some physically meaningful aspect of the line shape. The ideal statistic should be both robust and easy to interpret, and thus depends directly on the observations and on the science they are meant to support.

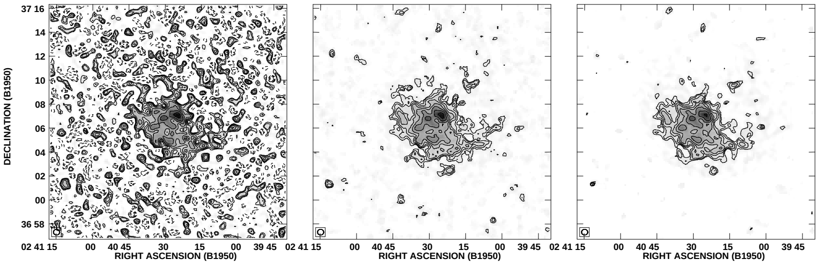


Figure 12–17. Moment 0 (total intensity) H I maps of the low-velocity channels in NGC 1058, illustrating the importance of clipping for even the lowest-order moment when the signal-to-noise ratio is low. The range of the grayscale and the contouring ($\pm 2^{n/2}\sigma$, with σ the noise in the left-most panel, and $n = 0, 1, 2, \ldots$) are the same in all three plots. The **left panel** was obtained by summing without clipping; the **middle panel**, by clipping the original cube at 2σ; and the **right panel**, by clipping the original cube wherever a smoothed (factor two in space and in frequency) version was below 3σ, with two channels 'spillover' to account for neighboring low-level emission.

11.2. Detection Experiments

So far this discussion has concentrated on the characterization of detected emission. What if the purpose of the observation is to find out whether any emission is present? This has been discussed elsewhere for continuum experiments (Lecture 9); spectral line work is different only in that the added frequency dimension complicates the search for emission. The key point is that a 3σ point in one channel becomes much more significant if the same or adjacent pixels are, say, 2σ in neighboring channels. In the past this has meant that the best way to find faint emission is to stare hypnotically at spectral line movies, letting the pattern recognition built into your eye and brain pick out noise spikes which appear to move in a fashion characteristic of, say, a rotating galaxy. A more objective method proposed by Juan Uson (and implemented in the AIPS task `SERCH`) is to smooth the cube with a variety of spatial and spectral filters, and check for significant emission at each resolution by calculating the probability of an observed peak given the (known) number of independent samples in the cube. In any case it is clearly an advantage to oversample the expected line width — if the source is (even marginally) resolved, the apparent motion of the signal from channel to channel is a strong diagnostic that would be lost with the use of coarser frequency resolution.

11.3. Beam Smearing

Beam smearing refers to the effects of observing an extended source with finite spatial resolution, and has important implications for the interpretation of observed line profiles. Consider the derivation of a rotation curve from observations of a disk in circular rotation. If the disk were infinitely cold, geometrically thin, and somewhat inclined to the observer, and if the spatial and spectral resolution were infinite, the problem would be trivial: simply read off the maximum Doppler shift along each line-of-sight, and that's the projected rotation speed. Now observe the same disk with poorer spatial resolution. Several spectra will

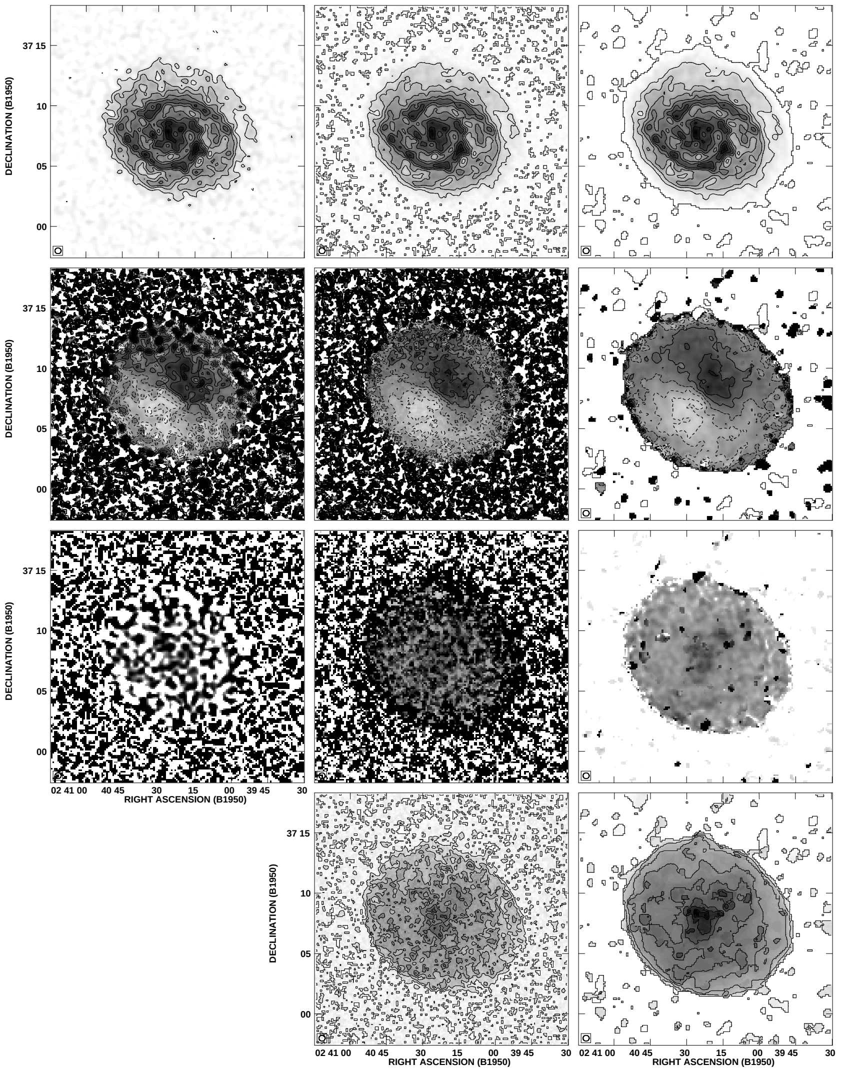


Figure 12–18. Moment maps of the H I in NGC 1058, illustrating the undesirable features of such analysis, even when the lines are simple, singly-peaked, and very strong (these are the same data shown in Figures 12–6 and 12–7). In the **left-hand column**, all the data were included; in the **middle column**, the cube was clipped at 2σ; and in the **right-hand column**, the cube was blanked where a version smoothed down by a factor two (both spatially and in frequency) was below 3σ, with an additional two-channel 'spillover' to allow for regions of fainter emission. The **top row** shows the moment 0 (total intensity) images, with linear contours in units of $N_H = 10^{20}\,\mathrm{cm}^{-2}$. The **second row** from the top shows the moment 1 (mean velocity) images, contoured at −20, −15, −10, ..., 20 km/sec. The **third row** from the top presents the second moments (velocity dispersions), the grayscale ranging from 0 to 20 km/sec. Finally, the **bottom row** shows the number of channels used in the moment calculation, contoured at 5, 10, 15, ..., 45 channels (the left-hand column is not plotted, as it is by definition constant, at 108 channels). The lower noise level in the middle and right-hand columns is due entirely to the exclusion of many channels which contribute noise but little or no signal.

be piled together, and the observed line profile will be the intensity-weighted sum of those spectra. Depending on the distribution of gas within the beam, the inferred rotation speed may be either higher or lower than the true value at the center of the beam. In the inner parts of a spiral galaxy, where the rotation curve is rapidly rising and the amount of gas often declining with radius, the effect is to systematically underestimate the true rotation speed, and hence infer a less rapid rise to maximum. The resulting estimates for the mass distribution may be seriously in error if this is not taken into account (see Begeman 1987, 1989 for a detailed discussion). Similarly one might easily infer an increase in the velocity dispersion towards the center of a spiral galaxy, simply because of the combination of finite spatial resolution with a rapidly-changing rotation speed[14]. Notice that this effect has nothing to do with poor *spectral* resolution — infinitely narrow channels would not help. The ultimate example of course is single-dish profiles, where one beam covers an entire galaxy; simply using the peaks of those profiles would systematically underestimate the true rotation speed (see Rhee & van Albada 1996, and references therein, for detailed discussion of this and other complexities in interpreting low-resolution data). Not everyone observes rotating disks, but while the specifics vary the basic problem is universal. For instance, line observations of molecular clouds with poor resolution cannot distinguish internal motions within clumps from the bulk motion of one clump relative to another. The point is simply that one must take account of structures on size scales below the observing beam when interpreting those observations. One can even turn this to advantage, using slight changes in position with velocity to image structures below the interferometric resolution; see Scoville, Yun, & Bryant (1997) for a nifty example (Arp 220).

11.4. Lines Which Track the Continuum

Often spectral lines can physically occur only in certain special locations. Most obviously, absorption and stimulated emission require a background continuum source, and in fact one is often interested in the line-to-continuum ratio rather than in the line intensity itself. This has implications throughout the reduction process. Bandpass correction and continuum subtraction are often particularly important, since one is searching for what may be a faint line directly atop a strong continuum source — any error can create apparent line emission/absorption which can easily mimic or obscure the true line profile. On a more positive note, a strong continuum source can be very useful for self-calibration, and the continuum may provide a very useful constraint for deconvolution algorithms, placing a limit both on the extent of the line and on its possible intensity. In the analysis stage one must remember that the noise in the line may be proportional to or correlated with the continuum signal, and that noise levels which are linear in flux density are *not* linear in optical depth. This is analogous to the case of polarization measurements (cf. Lecture 6).

[14] Careful analysis of spectra of face-on spirals indicates that the intrinsic velocity dispersion does indeed increase near the centers of spirals, making the situation even more confusing.

11.5. Comparing Different Lines

Another thorny problem is the accurate measurement of line ratios. The most important consideration here is that the (u, v) coverage be as similar as possible in the two transitions. An interferometer is a spatial filter, and identical spatial distributions observed with different baselines may appear wildly discrepant. Identical coverage may however be difficult to achieve, since one often wishes to compare lines widely separated in frequency, e.g., the rotational transitions of CO at 115 and 230 GHz. This together with intrinsic differences in the emitting region make careful deconvolution vital; the data should be tapered to make the dirty beams as similar as possible, and the identical restoring beam used for both lines. Proper calibration is also more important than for a single transition, since calibration errors can produce spurious structure as well as offsets in the line ratios.

12. Outstanding Problems

There are a number of obvious, important problems specific to line observations, for which the solutions are either unknown, incomplete, or not yet implemented. This in some ways is encouraging — there is still room for significant improvements in our techniques, which should lead to significant advances in our understanding. The following is a brief, incomplete, and highly subjective list of what seem to me among the most rewarding areas for further study.

- **Automatic flagging:** Compared to single-dish observations and data reduction at other wavelengths, flagging of interferometric data is both primitive and amazingly labor-intensive. This will become a more and more serious problem as instruments move to longer wavelengths and wider bandwidths, encountering an ever-increasing variety of radio frequency interference. Even trivial algorithms (see for instance the AIPS task `FLGIT`) can save enormous amounts of time, and rescue observations which would otherwise be effectively impossible.

- **The deconvolution of faint, extended sources:** One often wishes to compare line emission in different channels (or in different transitions) which have wildly different signal strengths and distributions. Ideally one wants a deconvolver which works equally well at both high and low signal-to-noise ratios, and for point-source and extended emission; which returns an estimate of the local resolution and SNR as well as a simple model; which preserves the character of the noise, rather than introducing very 'spikey' structures or very different behavior on (*vs.* off) the source; and which can be run effectively with minimal interactive guidance from the user. This sounds like a tall order but there are several promising approaches, e.g., the fractal pixon method of Piña & Puetter (1992, 1993; see also Dixon *et al.* 1996, 1997); and again even simple things like automatic boxing would be a big help.

- **Thinking in three dimensions:** Most current software thinks of spectral line data as a trivial extension of continuum processing. Allowing for 3D

> intelligence could lead to very large gains. For instance, a deconvolver could easily be designed which would use the fact that most real emission occurs in more than one channel, allow for the spectral response of the instrument rather than treating each channel independently, or require (*à la* MEM) that the emission change minimally between channels. Similarly one could imagine keeping track of the location of continuum sources, both to confine regions of absorption or stimulated emission, or to allow for increased noise or systematic errors there. Multiple transitions could be fit – or even deconvolved – all at once, rather than independently; the stacking of radio recombination lines for sensitivity is a current example of this approach, which could profitably be made more general. Self-calibration using all the line data is another natural extension, in this case of ongoing maser work.

If this is something of a wish-list for current observatories, there are other problems which absolutely must be solved before improved or new instruments can be used effectively. Wide bandwidths and huge numbers of channels are coming everywhere, forcing continuum observations to be a special case of spectral line work rather than the reverse, and mandating much more streamlined and automatic reductions. The Millimeter Array will require simple, reliable, and consistent mosaicing, with single dish data added in and with multiple transitions the norm rather than the exception. On longer time scales, the Square Kilometer Array will eventually produce unheard-of spectral dynamic ranges, requiring careful calibration and exquisite continuum subtraction. Both the challenges and the potential gains are tremendous.

Acknowledgments. I have benefited greatly from innumerable discussions with many of my colleagues at NRAO, particularly Jacqueline van Gorkom, Juan Uson, K.S. Anantharamaiah, Athol Kemball, Tim Cornwell, Ketan Desai, and Vivek Dhawan. Debra Shepherd and Aditya Dayal provided some of the images used in the examples. Dick Thompson pointed out an early error in the discussion of DSB systems. Finally, Andreea Petric, Marc Verheijen, and Min Su Yun provided useful comments on the manuscript.

References

The AIPS Cookbook 1998, ed. E. Greisen (available at `http://www.cv.nrao.edu/aips/cook.html`).

Begeman, A. 1987, *HI Rotation Curves of Spiral Galaxies*, Ph. D. Thesis, University of Groningen.

Begeman, A. 1989, *A&A*, 223, 47–60.

Bos, A. 1984, in *Indirect Imaging* ed. J. A. Roberts (New York: Columbia University press), 239–243.

Bos, A. 1985, *On Instrumental Effects in Spectral Line Synthesis Observations*, Ph. D. Thesis, University of Groningen.

Bosma, A. 1978, *The Distribution and Kinematics of Neutral Hydrogen in Spiral Galaxies of Various Morphological Types*, Ph. D. Thesis, University of Groningen.

Bosma, A. 1981a, *AJ*, 86, 1791–1824.

Bosma, A. 1981b, *AJ*, 86, 1825–1846.

Braun, R. 1995, *A&AS*, 114, 409–438.

Carilli, C. 1991, *Spectral Dynamic Range at the VLA: The 3 MHz Ripple* (VLA Test Memorandum 158).

Condon, J. J., Cotton, W. D., Greisen, E. W., Yin, Q. F., Perley, R. A., Taylor, G. B., & Broderick, J. J. 1998, *AJ*, 115, 1693–1716 (the NVSS survey). See also `http://www.cv.nrao.edu/~jcondon/nvss.html`

Cornwell, T. J. 1983, *A&A*, 121, 281–285.

Cornwell, T. J., Holdaway, M. A., & Uson, J. M. 1993, *A&A*, 271, 697–713.

Cornwell, T. J., Uson, J. M., & Haddad, N. 1992, *A&A*, 258, 583-590.

Dixon, D. D., Johnson, W. N., Kurfess, J. D., Piña, R. K., Puetter, R. C., Purcell, W. R., Tuemer, T. O., Wheaton, W. A., & Zych, A. D. 1996, *A&AS*, 120, 683–686.

Dixon, D. D., Tuemer, T. O., Zych, A. D., Cheng, L. X., Johnson, W. N., Kurfess, J. D., Piña, R. K., Puetter, R. C., Purcell, W. R., & Wheaton, W. A. 1997, *ApJ*, 484, 891–899.

Ekers, R. D. & van Gorkom, J. H. 1984, in *Indirect Imaging* ed. J. A. Roberts (New York: Columbia University press), 21–32.

Fisher, R. 1996, *A Program for Minimization of the Effects of Interference on GBT Observations* (GBT internal memorandum, available at `http://www.gb.nrao.edu/ rfisher/Interference/program.gbt`).

van Gorkom, J. H. & Ekers, R. D. 1989, in *Synthesis Imaging in Radio Astronomy* eds. R. A. Perley, F. R. Schwab, & A. H. Bridle (San Francisco: PASP), 341–353.

van Gorkom, J. H., Bahcall, J. N., Jannuzi, B. T., & Schneider, D. P. 1993 *AJ*, 106, 2213-2217.

Harris, A. I., Avery, L. W., Schuster, K.-F., Tacconi, L. J., & Genzel, R. 1995, *ApJ*, 446, L85–88.

Harris, F. J. 1978, *Proc. IEEE*, 51–83.

Herrnstein, J. R., Moran, J. M., Greenhill, L. J., Diamond, P. J., Miyoshi, M., Nakai, N., & Inoue, M. 1997, *ApJ*, 475, L17–L20.

Jörsäter, S. & van Moorsel, G. A. 1995, *AJ*, 110, 2037–2066.

Killeen, N. 1993, *Analysis of ATCA Data with AIPS (ATNF manual. No. K2)* (available at `http://www.atnf.csiro.au/computing/software/atca_aips/atcal_html.html`).

van der Kruit, P. C. & Allen, R. J. 1978, *ARA&A*, 16, 103–139.

van Langevelde, H. J., & Cotton, W. D. 1990, *A&A*, 239, L5.

Malhotra, S. 1994, *ApJ*, 433, 687–704.

Malhotra, S. 1995, *ApJ*, 448, 138–148.

Lillie, P. 1994, *Why the "3 MHz" Ripple Moves With Time* (VLA Test Memorandum 190).

Olling, R. P. 1995, *AJ*, 110, 591–612.

Olling, R. P. 1996a, *AJ*, 112, 457–480. Olling, R. P. 1996b, *AJ*, 112, 481–490.

Piña, R. K. & Puetter, R. C. 1992, *PASP*, 104, 1096–1103.

Piña, R. K. & Puetter, R. C. 1993, *PASP*, 105, 630–637.

Reid, M. J. & Menten, K. M. 1990, *ApJ*, 360,L51–54.

Rhee, M.-H. & van Albada, T. S. 1996, *A&AS*, 115, 407–437.

Rupen, M. P. 1997, *VLA Scientific Memorandum No. 172: A Test of the CS (Shortened C) Configuration* (available at `http://info.aoc.nrao.edu/doc/vla/html/Memos/scimemolist.shtml`).

Sault, R. J. 1994, *A&AS*, 107, 55–69.

Sault, R. J. 1995, *A&AS*, 109, 593–595.

Sault, R. J. & Killeen, N. 1998, *Miriad Users Guide* (available at `http://www.atnf.csiro.au/computing/software/miriad/userhtml.html`).

Schilke, P., Groesbeck, T. D., Blake, G. A., & Phillips, T. G. 1997, *ApJS*, 108, 301–338.

Steer, D. G., Dewdney, P. E., & Ito, M. R. 1984, *A&A*, 137, 159–165.

Thompson, A. R., Moran, J. M., & Swenson, G. W. 1991 (TMS), *Interferometry and Synthesis in Radio Astronomy* (Malabar, Florida: Krieger Publishing Company).

Torrelles, J. M., Gómez, J. F., Rodríguez, L. F., Curiel, S., Ho, P. T. P., & Garay, G. 1996, *ApJ*, 457, L107–L111.

Torrelles, J. M., Gómez, J. F., Rodríguez, L. F., Ho, P. T. P., Curiel, S., & Vázquez, R. 1997, *ApJ*, 489, 744–752.

Verheijen, M. A. W. 1997, *The Ursa Major Cluster of Galaxies: TF-Relations and Dark Matter*, Ph. D. Thesis, University of Groningen.

Wilcots, E. M., Brinks, E., & Higdon, J. 1995, *A Guide for VLA Spectral Line Observers* (VLA internal document) (available at `http://www.aoc.nrao.edu/doc/vla/html/specline.shtml`).

Wakker, B. P. & Schwarz, U. J. 1988, *A&A*, 200, 312–322.

Wright, M. 1991, *Optimized Calibration* (BIMA Memoranda No. 19, available through `http://bima.astro.umd.edu/bima/memo/memo.html`).

Wright, M. 1994, *Passband Correction with the BIMA Array* (BIMA Memoranda No. 33, available through `http://bima.astro.umd.edu/bima/memo/memo.html`).

Yun, M. S., Ho, P. T. P., & Lo, K. Y. 1994, *Nature*, 372, 530.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

13. High Dynamic Range Imaging

Richard A. Perley
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract.
The origin of errors that affect image quality are discussed, along with methods to reduce such errors and increase dynamic range in the final image.

1. Introduction

In recent years, outstanding images of the radio sky have been produced from interferometric data obtained with modern, high-precision synthesis arrays such as the VLA, the Netherlands' Westerbork Synthesis Radio Telescope, and the U.K.'s MERLIN. Production of these images is by no means automatic, for the data are invariably corrupted by a host of errors, due to both atmospheric and instrumental effects. Removal of these errors is now possible to a degree unimaginable a few years ago, thanks to sophisticated new algorithms which, when employed by a cognizant user, allow accurate imaging, often with the resulting image noise near the theoretical limit. The complexity of these algorithms, especially in regard to their interaction with the user, requires familiarity with their use and effect. It is important to recognize from the outset that these processes do *not* constitute a 'black art'. The stunning results recently achieved come from careful application of simple and basic principles to interferometric data. The purposes of this lecture are to identify and discuss the origin and effect of some (but certainly not all) of the errors that can affect interferometric data, and to discuss and demonstrate the application of modern techniques and data-processing algorithms to correct these errors.

This lecture is organized as follows: In Section 2, I briefly discuss the differences between two measures of image correctness: Image Fidelity and Image Dynamic Range. Following this, in Section 3, I show how errors in visibility amplitude and phase affect an output image. The origin of errors which affect image quality is discussed in Section 4, and Section 5 demonstrates the methods that are useful in improving an image, using data taken by the VLA as an example.

2. Image Fidelity *vs.* Image Dynamic Range

In radio astronomical imaging, one accumulates a large body of information (the visibilities), from which is produced an estimate of the sky brightness. We are, of course, primarily interested in getting the right answer. However, since the data are generally corrupted, and since not all the desired measurements of the visibility are present, the output image must contain errors. Although the processes described elsewhere in this book (such as self-calibration and deconvolution) can be counted on to reduce these errors, there will always remain small, but often important deviations from the correct image. The term 'Image Fidelity' is here defined as the difference between any produced image and the correct image. The problem is to determine this quantity. This is a difficult task since one doesn't generally have *a priori* knowledge of the correct image. It might seem possible to compute the error in any given pixel of an image through knowledge of the deconvolution process combined with the (u, v) plane coverage, the source structure, and distribution of known errors amongst the baselines. Indeed, it is clearly desirable that any image be accompanied by an image describing the likely error of every pixel in it. Unfortunately, I am told by knowledgeable people that this computation, even if it were theoretically possible, is well beyond our current capabilities.

Yet, it is clearly desirable to have some quantitative estimate of the errors in an output image. Since no quantitative measure has emerged to allow calculation, or estimation of the 'fidelity', it has become common to use 'dynamic range', which has the singular merit of having a simple definition. Unfortunately, the connection between the dynamic range and image fidelity is hard to make.

Dynamic range is widely defined as the ratio between the peak brightness on the image and the r.m.s. noise in a region believed to be void of emission (such regions, fortunately, are commonplace in astronomy). High dynamic ranges imply low errors, so it is thus implied that dynamic range is a measure of the accuracy of the resultant image. This can be misleading. What is true is that the noise in an empty region represents an easily calculable lower limit to the error in the brightness of a non-empty region. The true error distribution is non-uniform; indeed, the errors of calibration must result in errors in the image which behave somewhat like the sidelobes in the beam, since they are constrained to the same (u, v) cells sampled by the data. These sidelobes are almost always greatest near the peak, so that, in an image, the errors will be greater in regions containing structure. Furthermore, errors in deconvolution must be considered. These also are spatially variant, with the net effect of increasing the probable error in the reconstructed brightness of a non-void region.

Unfortunately, there is no general guideline for estimating the effects of errors caused by use of deconvolution. It is perhaps ironic that the techniques upon which we have become so dependent, and which are unquestionably improving our product, leave us with a product whose accuracy cannot be estimated quantitatively. All we can say for sure are: (a) that the deconvolved image is much closer to the real sky than the 'dirty' image, and (b) that a lower limit to the error in the deconvolution is given by the noise in the surrounding void regions.

However, I will now argue that despite the firm theoretical understanding of the errors accompanying a deconvolution, the noise in a blank region of an image made with sufficiently well-calibrated data adequately covering the (u, v) plane will be a good indicator of the true error. Under these conditions, Dynamic Range will be a good indicator of Image Fidelity, in the sense that improving the former will also improve the latter. Self-calibration of visibility data from simple, strong, isolated objects reduces the 'noise' in regions of no known structure and increases the source brightness. That is, the 'dynamic range', as defined above, increases. This increase is invariably accompanied by reduction, or disappearance, of features known (or suspected) to be false. These features are usually of a non-physical nature, such as parallel ridges of positive and negative amplitude which cover much of the image. These changes agree with our intuition concerning the appearance of the radio sky—so it is believed that the improvement in dynamic range which usually accompanies self-calibration does represent an increase in image fidelity. As the process of self-calibration proceeds, physically reasonable structures (such as background objects and regions of low surface-brightness) which were formerly masked by errors become clearly discernible. All of this is as it should be, according to our intuition. These encouraging indications are not isolated to simple point sources, as the same process occurs for large objects, provided the (u, v) coverage is adequate. We can judge this by seeing if the resulting noise is 'flat'—i.e., spatially homogeneous. If important measurements of the visibility are missing (important here means changes in the visibility, representing structures of larger scale-size) the resulting effect on an image is spatially variant residuals—waves or bowls of enhanced residuals. One can be fairly certain that all important regions of the (u, v) plane are sampled, and the data well calibrated, if the resulting noise is flat. I shall show an example of inadequate coverage later in this lecture.

Thus, dynamic range shall be employed in this lecture as the parameter with which to judge the fidelity of the images which result from processes discussed here at length. The reader must keep in mind that the noise estimates determined in this manner will be a lower limit to the true, position-dependent errors in the image.

3. The Effects of Visibility Errors on Image Dynamic Range

The dynamic range attained on an image will depend, amongst other things, on the type, size and distribution of errors in the measurements of the visibility. In this section, I apply some simple arguments to allow rough calculation of the dynamic range, given baseline-based or antenna-based errors of various magnitudes. For the purpose of analysis, it is simplest to consider a point source—for which all visibilities are the same. The analysis will be done in one dimension, but the extension to two is straightforward.

Consider a single 'snapshot' observation[1] of a unit amplitude source located at the phase-tracking center, using N antennas. Assuming all correlations are made, there are $N(N-1)/2$ complex visibilities. Suppose all but one are perfect—i.e., they have unit amplitude and zero phase. Their visibilities are described by $V(u) = \delta(u - u_k)$, while the discrepant visibility (from a baseline of length u_0) is

$$V(u) = \delta(u - u_0)e^{-i\phi}\,, \tag{13–1}$$

where ϕ is the phase error (in radians), and δ is the Dirac delta function. The image is formed by evaluating the transform $I(l) = \int V(u)e^{i2\pi ul}\,du$, so for each 'good' baseline, the integral gives a contribution of $2\cos(2\pi u_k l)$. (The factor of two arises because each visibility is counted twice, once at the position u_k, and again, with its complex conjugate, at $u = -u_k$.) The 'bad' baseline contributes $2\cos(2\pi u_0 l - \phi)$, which for small ϕ becomes $2[\cos(2\pi u_0 l) + \phi \sin(2\pi u_0 l)]$, so that the resulting image is

$$I(l) = 2\phi \sin(2\pi u_0 l) + 2 \sum_{k=1}^{N(N-1)/2} \cos(2\pi u_k l)\,, \tag{13–2}$$

[1] A snapshot is a single short (< 10 min) observation with a two-dimensional array.

while the beam, or point spread function is

$$B(l) = 2 \sum_{k=1}^{N(N-1)/2} \cos(2\pi u_k l)\,. \tag{13–3}$$

Defined in this way, and with a quasi-uniform distribution of spacings, the beam and image both have amplitude $N(N-1)$, and width $\sim 1/u_m$ radians, where u_m is the maximum spacing (in wavelengths). Deconvolution in this case is accomplished by subtracting the beam from the image, giving a residual, $R(l) = 2\phi \sin(2\pi u_0 l)$, a periodic function of amplitude 2ϕ and period $1/u_0$. Note that the phase error results in an *odd* residual (as required by the arguments in Lecture 15), whose amplitude is proportional to the error (for small errors). If the Dynamic Range, D, is defined as $D =$ (peak on image)/(r.m.s. on image residual), then

$$D = \frac{N(N-1)}{\sqrt{2}\,\phi} \approx \frac{N^2}{\sqrt{2}\,\phi}\,, \tag{13–4}$$

with the approximation valid for large N.

Analysis of an amplitude error is similar. In this case, write the visibility of the 'bad' baseline as $V(u) = (1+\epsilon)\delta(u-u_0)$. Following through, the same results as before are recovered via the substitutions

$$\phi \to \epsilon \quad \text{and} \quad \sin \to \cos\;. \tag{13–5}$$

An important conclusion from this exercise is the following:

A 10° phase error is as bad as a 20% amplitude error.

It is a well-known fact that, in the process of self-calibration of VLA data, the phase correction has a much greater effect on the image dynamic range than does the amplitude correction. The statement in the box explains why. The effect of the atmosphere upon the visibilities is almost entirely in phase—ten-degree phase errors are commonplace, while the corresponding amplitude error (20%) is extremely rare (and, if present, due to instrumental problems). In contrast, many types of instrumental errors do not discriminate between amplitude and phase—these errors affect the Cos or Sin correlators independently, with no particular relation between them. For these cases, the amplitude and phase errors, when expressed as percentage loss (or gain) in amplitude, and degrees of phase, will be about equal.

The extension of these simple arguments to the case where there are many random errors on many correlators or antennas is straightforward. Suppose each baseline has a random error typically of the magnitude given above. Then, the dynamic range must be decreased from the single correlator case by a factor $\sqrt{N(N-1)/2}$, giving

$$D = \frac{\sqrt{N(N-1)}}{\phi} \approx \frac{N}{\phi}\,. \tag{13–6}$$

Suppose now that the error is antenna-based, and that only one antenna has an error. Thus, instead of one bad baseline, there are $N-1$. Then, again assuming incoherence in the noise (which is approximately right), the dynamic range becomes

$$D = \sqrt{\frac{N-1}{2}}\,\frac{N}{\phi} \approx \sqrt{\frac{N}{2}}\,\frac{N}{\phi}\,. \tag{13–7}$$

If all antennas have a typical phase error of this magnitude, the dynamic range will decrease by $\sqrt{N}$, yielding

$$D = \frac{1}{\phi}\sqrt{\frac{N(N-1)}{2}} \approx \frac{N}{\sqrt{2}\,\phi}\,. \tag{13–8}$$

This result differs by a factor of $\sqrt{2}$ from the result derived by assuming an independent error for each correlator, since the effect on any baseline in the antenna-based case is the quadratic sum of the errors on each of the contributing antennas.

Thus, for any given snapshot, the effect of typical errors of order ϕ is to give a dynamic range limit of order N/ϕ. The effect of multiple snapshots is simple as long as the errors are different between snapshots. Here, 'different' means the error of a typical (u, v) measurement has changed by an amount approximately equal to the error itself. This can arise either from a change of the error (such as given by a changing atmosphere), or by rotation of the baseline coordinate, since this changes the spatial frequency and (more importantly, for two dimensions), rotates the fringes on the image, so that the particular (u, v) component is measured by a different baseline, with a different error. If M successive snapshots are independent, then the dynamic range becomes

$$D = \frac{\sqrt{M}\sqrt{N(N-1)}}{\phi} \approx \frac{\sqrt{M}N}{\phi}\,. \tag{13–9}$$

The difficulty in using this formula for continuous observing is that there is no clear time scale over which the errors can be considered 'independent'. The atmospheric time scale for significant phase changes is anywhere from one-half minute to a couple of hours, while the typical time for significant rotation of the baseline might be a few seconds to tens of minutes, depending on baseline length and field-of-view. Thus, M can range from about 25 to a few thousand for a twelve-hour observation.

These equations, though simplistic, return reasonable results. They predict that for a single snapshot with 27 antennas, residual calibration errors of order ten degrees will limit the dynamic range to approximately 1500 : 1 if confined to a single baseline, 700 : 1 if due to an antenna, and 100 : 1 if equally distributed amongst all antennas.

These expressions can be used to allow us to estimate the magnitude of residual errors. At the VLA, it is found that after regular calibrations (i.e., using an external calibrator), the typical best dynamic range for an observation will be about 1000 : 1. Assuming fifty independent observations, we find the residual error is 0.2 radians, or ten degrees. After self-calibration, the typical maximum dynamic range will be $\sim$ 20,000 : 1, indicating residual errors of 0.01 radian, or 0.6 degrees. These are the levels expected to arise from 'closure effects'—errors that are dependent on baselines, rather than antennas, and which therefore cannot be removed by self-calibration. It is found that, in the best cases, correction for closure errors results in dynamic ranges of $\sim$ 80,000 : 1, corresponding to errors of 0.003 radians, or 0.2 degrees. This level probably represents the time-variable part of the closure errors. Finally, use of the VLA spectral line correlator allows one to avoid closure errors to a large degree. Using this system, the best dynamic range yet achieved is about 350,000 : 1, corresponding to residual errors of 0.0005 radians, or 0.04 degrees. The next problem is to identify the origin of this remaining error. This is attempted in the next section.

These equations can also be used to estimate the required phase accuracy needed to attain the theoretically best dynamic range. Anticipating a result derived in the next section, the maximum potential dynamic range attainable with the VLA is of order $10^7:1$. Reaching this level will require the visibilities to be accurately measured to about one part in 10^5, corresponding to a phase accuracy of better than 10^{-3} degrees.

Another interesting application of these ideas is to the problem of the required accuracy of baseline determinations. In Lecture 5 it was pointed out that the phase errors that arise from observing with an erroneous baseline determination will largely be removed through calibration by a nearby point source. However, the more permanent legacy of an erroneous baseline will be in the image, since the (correctly) calibrated data will be gridded in the wrong place, resulting in an error. The preceding analysis can be used to estimate the magnitude of the effect, when it is noted that an error in gridding a visibility point by an amount δu results in a phase error of $2\pi\delta u l_0$ to that visibility, where l_0 is the offset of the object in question from the phase-tracking center. It is then seen that the resulting dynamic range limitation is dependent on the position of the object in question. Since the error is in the position of an antenna, the resulting phase error will be distributed amongst all N antennas of the array, resulting in $N-1$ baselines with the same phase error. For a two-dimensional array, however, the orientation and separation of these $N-1$ spacings will be largely random, so use of Equation 13–6 is justified. From this equation, and inserting the above error, the following limit on the accuracy of the baseline is determined:

$$\delta u = \sqrt{\frac{N}{2}}\,\frac{N}{2\pi D l_0}\,. \tag{13–10}$$

Again taking the VLA for an example, the relation becomes $\delta u = 16/Dl_0$. At 20 cm, with a dynamic range of 1000 : 1, an accuracy of 1λ is indicated, or about 1 nsec. The situation is actually somewhat worse than indicated here, since our putative object is not the only one in the beam. At 20 cm there will be many such objects, each contributing a separate set of errors to the image, which may be assumed to add quadratically, so that the required accuracy just rises roughly with $\sqrt{n}$, the number of such objects. At lower frequencies, the increased angle calls for increased accuracy of baselines, but the similarly increased wavelength offsets this, so the necessary physical accuracy remains the same, perhaps a few centimeters. Fortunately, determining the VLA's baselines to this accuracy is routine.

4. Origin of Residual Errors for the VLA

The ultimate dynamic range will be that which is set by the thermal noise. This result, which is derived in Lecture 9, can be written as

$$\delta I = \frac{C}{\sqrt{N(N-1)\Delta\nu\tau}}\,, \tag{13–11}$$

where C is a constant depending upon the antenna size and efficiency, the system temperature, and the type of correlator. It is important that anyone contemplating enhancement of an image know the 'base' noise level. If the noise level on the current image is close to the theoretical limit, there is no more to be done. Note that the potential dynamic range for some objects is enormous. For example, consider the famous quasar 3C 273. Its core flux density is approximately 35 Jy, while the theoretical noise, for the VLA with 35 hours of observing at 6 cm, is about 6 μJy. Thus, the theoretical maximum dynamic range exceeds six million to one! This is of course an extreme example, and it might be argued that so few objects exist which can be imaged at this level that it is not worth considering what sources of error are important at this level. However, the current maximum dynamic range of order 350,000 : 1 can be reached (at least in principle) with any object whose brightness exceeds $\sim$ 1 Jy/beam. There are hundreds of such objects known whose (possible) low-level structure is hidden by the remaining errors.

In the following sub-sections, I will consider the factors which may be important in limiting dynamic range. I emphasize *may be*, since in most cases the necessary tests have not yet been done. The first three sub-sections below review the sources of error already mentioned.

4.1. Atmospheric phase errors

Atmospheric phase errors limit the dynamic range to values ranging from several to 1 to perhaps 1000 : 1, depending on resolution and atmospheric conditions. Since their effect is to modify the phase of an antenna, they can be corrected by self-calibration, provided that sufficient signal-to-noise ratio and a sufficiently good model exist. See Lecture 10 for a fuller discussion of these requirements, and Section 5 of this lecture for a demonstration of the techniques of using self-calibration.

4.2. (u, v) coverage errors

The process of deconvolution in essence interpolates visibility values for (u, v) spacings which were not measured by the interferometer. This is not an error-free process. Clearly, the fewer gaps in the coverage, the less likely it is that the deconvolved image will be in error. Different algorithms, e.g., 'CLEAN' or Maximum Entropy, will interpolate differently. The question of which does a better job is very complicated and has not been well studied, as I alluded to in Section 2. It should be noted that it is especially important to obtain measurements of the visibility in regions where it is changing rapidly, if all structures present are to be properly represented in the output image.

4.3. ‘Closure errors’

‘Closure errors’ imply errors in the measured visibility which are not factorable into products of antenna-based errors. Thus, they are unique to the baseline involved. They result from various sources; for the VLA, the most important appear to be errors in the antenna delay settings, mismatches in the antenna IF bandpasses (Clark 1978, Thompson 1980), and non-orthogonality between the Cos and Sin correlators (Bagri 1990), caused by the analog phase-shifter in the complex correlator. The first error can be minimized by making careful delay settings prior to observation. The second error is minimized by careful matching of the final passband filters. At the VLA, it appears that this causes amplitude closure errors at the 1% level, but that phase closure errors of a few degrees must have some other origin. The latter error causes the Cos and Sin fringe patterns to be non-orthogonal on the sky. It can easily be shown that this error is baseline-dependent, so it cannot be corrected by normal calibration. It appears that this is the most important origin of closure errors for the VLA’s continuum correlator.

All these effects can, in principle, be measured, and corrections applied to the affected data. Obviously, if the errors are time-independent, the process is much simpler. This procedure is generally referred to as ‘baseline calibration’. A better solution, for the VLA at least, is to use the spectral line correlator, which avoids the non-quadrature problem through use of a computed Hilbert transform. The delay setting error is eliminated since the channel widths are much more tolerant of this error, and the bandpass mismatching problem is eliminated by discarding those channels on the edges of the bandpass, where the errors arise. (Or, even better, a bandpass calibration can be performed which effectively resets the delays and matches the bandpasses in the same operation.) The cost of using a spectral line correlator is a greatly increased computing load and a narrower bandwidth and hence lower sensitivity.

The following sections describe possible origins of errors responsible for limiting the dynamic range to the current limit of approximately 100,000 : 1.

4.4. Quantization correction (‘Van Vleck correction’)

Jon Romney has shown in Lecture 4 that the correlation coefficient produced by a digital correlator differs in a nonlinear way from that produced by a perfect analog correlator. This difference is a complicated function of the true correlation coefficient and the signal powers, and it can be computed following, for example, the expressions given by Schwab (1979) and by D’Addario *et al.* (1984). Note that the required corrections must be applied independently to both Cosine and Sine correlations (in-phase and quad-phase), so the net effect on the complex visibility depends on the actual interferometer phase at the time of measurement.

In general, the corrections to the Cosine and Sine correlations are different, so the correction will affect both the amplitude and the phase of the observed visibility. In some special cases, however, the corrections affect the amplitude only. This situation occurs for all odd multiples of $\pi/4$ radians, since in this case the correlation in the two channels is equal, so the corrections are equal, and for all even multiples of $\pi/4$ radians, since in this case all the correlation is in one channel. It is a fact that the correction depends on the observed amplitude

Table 13–1. Errors due to Quantization for the 3-Level Correlator

True Correlation	Measured Correlation	Corrected Correlation	Percentage Error
.04	.0340	.0400	.008
.08	.0687	.0800	.040
.12	.1031	.1201	.095
.16	.1375	.1603	.171
.20	.1721	.2005	.27
.50	.4364	.5236	1.72
.80	.7188	.8378	4.72

and phase, and on the signal powers—and this is the strongest argument for applying the corrections on-line. By their very nature, source visibilities vary with baseline and time. Even for the simpler case of a strong point-source, the atmosphere alters the observed visibilities on a short time scale. (For strong point sources, the VLA allows a phased-array mode of observing, which puts essentially all the correlation into the Cos channel, eliminating the need for quantization correction, provided the coherence amplitude does not significantly change in time or with baseline. This method of avoiding the problem is effective only for point-dominated objects.) Correction of the visibilities after calibration is difficult or impossible, depending on whether the necessary information has been preserved or not. At the present time, these corrections are not applied to VLA data on-line, but can optionally be applied offline with reduced accuracy by the AIPS task 'FILLM'.

Note that since this error is baseline-based, it cannot be removed through antenna-based self-calibration. Furthermore, since the error is also time-variable (due to changing visibilities), it cannot be removed by the 'closure' correction techniques discussed in the next section.

We can get an idea of the magnitude of the error by using the expressions given by Schwab (1979) for the 3-level cross-correlator. For simplicity, suppose we observe a point source, and have arranged the array to auto-phase, so that all the correlated power is in one correlator channel (say, the Cosine channel). There then results a nonlinear relation between true and measured correlation amplitude, but no error in phase. The amplitude observed is in arbitrary units—I'll call them 'counts'. To convert these to physical units, 'Janskys', one must observe a source of known flux density. The subtle difficulty is that even if we auto-phase on this object also (making the phase error zero), the amplitude scaling we obtain will be different than for the object source, due to the nonlinearity of the error relation. Suppose we calibrate the interferometer by observing a point source whose true correlation is 0.02 (corresponding approximately to 10 Jy for the VLA) and applying the derived scaling to the target source. Table 13–1 shows the error in correlation for various values of true correlation. The first column lists the true correlation coefficient, the second a quantity representing the output of the 3-level correlator, the third that same output after calibration using the source of 0.02 correlation, and the fourth the relative amplitude error.

A 50 Jy source has a correlation coefficient of approximately 0.1, so the error in the visibility, after calibrating by a typical 2 Jy calibrator, amounts to approximately 0.06%. Although this analysis presumed that the correlation was confined to one correlator, the general situation, with correlation in both correlators, will give an error of the same magnitude. Thus we deduce that the typical phase error will be of order $0\overset{\circ}{.}03$. (Recall the boxed relation in Sec. 2.)

This simple analysis shows that for sources of modest strength (i.e., less than, say, 100 Jy) the error due to omitting the quantization corrections is very small. It is, nevertheless, large enough to limit the dynamic range to a level of a few hundred thousand to 1. For sources with higher coherence, the errors rapidly increase so that failing to correct for this effect should result in easily detected errors in the image.

4.5. Polarization leakage

Lecture 5 has discussed the calibration of astronomical data. But note that the polarizers are not perfect—each output contains some component of the opposite polarization. For example, if the voltage outputs of the right- and left-circularly polarized channels are denoted v_R and v_L respectively, we can write

$$\begin{aligned} v_R &= E_R e^{-i\chi} + D_R E_L e^{i\chi} \\ \text{and} \quad v_L &= E_L e^{i\chi} + D_L E_R e^{-i\chi} \,, \end{aligned} \tag{13–12}$$

where the E's are the circularly polarized signals which would be received by a perfect system, the D's are the 'leakage' terms, and χ is parallactic angle, accounting for rotation of the antenna on the sky. We note from this that the various products of these voltages are functions of all four Stokes parameters, instead of just two, which would occur if the leakage terms were identically zero. Assuming the circular polarization is zero, we can derive the following equations for the response of the parallel-hand correlators (the corresponding expressions for the crossed-hand correlators are given in Lecture 5),

$$\begin{aligned} V_{RR} &= V_I(1 + D_{R1}D^*_{R2}) + (V_Q + iV_U)D^*_{R2}e^{-2i\chi} + (V_Q - iV_U)D_{R1}e^{2i\chi} \\ \text{and} \quad V_{LL} &= V_I(1 + D_{L1}D^*_{L2}) + (V_Q + iV_U)D_{L1}e^{-2i\chi} + (V_Q - iV_U)D^*_{L2}e^{2i\chi} \,. \end{aligned} \tag{13–13}$$

The subscripts (1 and 2) refer to the two antennas comprising the baseline. The antenna feeds are assumed to be identical, and the V_I, V_Q, and V_U are the complex visibilities corresponding to the subscripted Stokes parameters. To assess the effects of leakage errors, first note the sizes of the various terms, taking $V_I = 1$ for reference. For nearly all sources of interest, the V_Q and V_U terms are less than 0.1 (i.e., less than 10% linear polarization), while the cross-polarization terms D range from 0.01 to 0.05. With these typical coefficients, the above equations show that the quantity V_{RR} (for example) is equal to the desired (and correct) V_I multiplied by a small, constant term $(1 + D_{R1}D^*_{R2})$, and further modified by the addition of two terms representing the leakage of linearly polarized flux into the desired correlation.

The multiplicative error is very small (typically 0.05%), time invariant (provided the antenna polarizations are also), baseline-dependent, and proportional to the desired correlation. It can thus be considered a 'closure error' and should be correctable by the techniques described in the next section. The additive errors are much more difficult to handle, since they are of appreciable magnitude (typically 1% to 0.1%), time variable (due to the change of parallactic angle with time), and not related to the desired quantity (V_I). Physically, the visibilities corresponding to the linearly polarized flux are being attenuated and phase-shifted by the leakage terms, rotated by the parallactic angle, then added to the visibility of the total intensity. Under these conditions, no coherent image of the polarized flux will result, and the net effect is 'rubbish' on the image, causing a reduction in the dynamic range. The expected amount of this leakage is at the level necessary to limit dynamic range at the observed limit—about 100,000 : 1. Note that this requires the source to be polarized, so, if this mechanism operates as proposed here, imaging of unpolarized sources should give higher dynamic range than imaging of polarized sources.

Finally, note that the effects of these leakages should be removable after correlation, since the equations show that the obtained correlations (V_{RR}, V_{LL}, V_{RL}, and V_{LR}) are linear combinations of the desired correlations (V_I, V_Q, V_U, V_V), with coefficients that are functions of χ and the leakage terms, D. The former of these is known (being determined by the geometry), while the latter is usually determined independently, through observations of strong point sources (see Lecture 5).

4.6. Loss due to phase winding

It should be obvious that coherent integration longer than the coherence time characterizing a phase instability will cause a loss of correlation. The causes of phase instability are multitudinous, and include tropospheric and ionospheric irregularities, effects of finite correction of phase rotation in the electronics, incorrect geometry of the interferometer, etc. For each of these, the results will be a loss in the measured coherence. To show the magnitude of the effect, I shall assume there is a constant linear phase-wind through the integration period. The complex response can then be expressed as

$$R = \frac{Ae^{i\phi_0}}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} e^{i\dot{\phi}t}\, dt\,, \tag{13–14}$$

where ϕ_0 is the phase at the center of the integration, and $\dot{\phi}$ is the phase-rate (rad/sec). The result of this integration shows the resultant phase is unaffected, but that the observed amplitude is reduced by a factor

$$\mathrm{sinc\ f\Delta t}\,, \tag{13–15}$$

where sinc is defined as sinc x = $(\sin \pi \mathrm{x})/\pi \mathrm{x}$, and $f = \dot{\phi}/2\pi$.

Note that a phase-wind of ten degrees across the integration window causes an amplitude loss of 0.1%, sufficient to limit dynamic range to a level of a few hundred thousand to one. A good observing strategy for those interested in these dynamic ranges is to choose the observing time to prevent atmospheric or ionospheric phase-winds of less than, say, five degrees. Experience shows that at the VLA, on an average summer day, the phase-wind on the longest baselines due to the atmosphere is typically 0.3 degrees per second per gigahertz. The rate under good conditions is perhaps one-tenth of this value.

4.7. Noise introduced through self-calibration

The visibilities are measured with finite signal-to-noise ratio, and some error is to be expected in the solution for antenna phase and amplitude. The self-calibration lecture shows that the expected error for the solved amplitude or phase is approximately

$$\sigma_G = \frac{\sigma_V}{S\sqrt{N}}\,, \tag{13–16}$$

where I have assumed a large number of antennas. S is the source flux density, σ_G is the error in the gain, and σ_V^2 is the variance of the receiver noise on one baseline (in the same units at S), for the time interval for which corrections are deemed necessary. Taking, for an example, a VLA observation of a 1 Jy point source, with 27 antennas and 10 second integration with 50 MHz bandwidth at 6 cm, the typical solution error will be $\sim 0.4\%$, sufficient to limit dynamic range to a few tens of thousands to one.

Improving the solution accuracy requires a longer solution time. However, note that this time cannot be extended indefinitely. The solved phase is applied with a linear interpolation. If the true atmospheric phase deviates appreciably from the interpolated value within the solution time, then a loss of dynamic range will result.

4.8. Effects of digital representation of models

In order to apply self-calibration we must represent the source of radiation, which in general is smooth and continuous, by a set of point components, or 'delta' functions. What is the error inherent in this procedure? For simplicity, I will discuss only the point-source case. There is no conceptual problem with modeling a point source with a single delta function when that object lies at the center of an image cell. But what if it does not? Suppose the object lies between two cells. Obviously, more than one component is required. In fact, it is easy to show that, in the general case, an infinite number of components is required, of both positive and negative sign. Truncation of this series results in an error between the model and the data, which will in turn lead to an error in any self-calibration solution, and ultimately limit the fidelity. Unfortunately, there is a commonly held 'rule' that, in self-calibration, one must not include negative components since they are 'unphysical'. This is true if a point source lies at a cell center, but is false if it does not. You can understand the necessity of negative components by considering the situation of a point object situated exactly between two cells. It is obvious that the two largest 'CLEAN' components will be positive, and in the two adjacent cells. But using just these as the model tells the self-calibration program that the source is double, with separation equal to the cell separation. Self-calibration with this model will result in antenna gains on distant antennas being lowered at the expense of antennas near the center. The purpose of the negative components is to 'eat away' at the double, making it look more like a point. As I stated, it can be shown in this case that a representation by an infinite series of points of alternating sign results in no error. Since we must truncate our models with finite numbers of components (not to mention the effect of noise or errors in visibility), one must ask how many components are needed for an acceptable model.

Table 13–2. Maximum Percentage Amplitude Error

	Number of Components				
Pts./beam	2	4	6	8	10
1.5	43	17	9	5	5
2.5	23	13	3	1.5	1.5
4	6	3	2	—	—
7	2	1	—	—	—

To investigate this question, I generated a database mimicking a full twelve-hour VLA observation of a 1 Jy point source. Phase shifting these data by 1/4 and 1/2 cell, followed by 'CLEAN' deconvolution, resulted in models with both positive and negative components, as described above. I then applied self-calibration with various numbers of components, for various cell-sizes (i.e., varying the number of points per beam). Table 13–2 summarizes the results.

Shown are the *maximum* amplitude errors in the self-cal solution for the given combination of number of components and number of points per beam. In all cases, this worst case occurs on the end antennas of the arms. The typical error is about one-tenth of the maximum error.

Thus, the first row represents the critically sampled case. The second row represents the 'typical' recommended sampling. The bottom two rows represent greatly oversampled beams. For this example, the source was shifted to half-way between two adjacent cells, so an even number of components was always used. Generally, the first two components are positive, the next pair negative, and so on. (This arrangement breaks down when heavy oversampling is used—but, in this case, virtually all the flux is in the central pair of cells, so that the contribution from adjacent cells is unimportant.) Note that simple two-component (all-positive) models result in very large errors when the sampling is near-critical. The table shows there are two ways to reduce this error: (i) use both positive and negative components, or (ii) over-sample the beam by at least a factor of two (i.e., at least 5 points per beam). Of these, the latter seems to return better solutions (although the slow improvement of the solution with increasing number of components may be due to a curious computation error discovered during the course of this investigation, the cause of which has not yet been discovered). The problem with solution (i) is that it will be hard to separate the physical negative components which are required by sampling from those unphysical components due to bad calibration. The problem with (ii) is simply one of computing. Doubling or quadrupling the number of points per beam means a concomitant increase in the image size.

There is a third method which might be quite attractive. This is fractional-cell deconvolution. Instead of limiting the beam subtraction to cell centers, this method would interpolate to find the maximum, and then subtract from the cell locations an interpolated beam. This technique would greatly reduce the number of components, and allow near-critical sampling. However, deconvolution methods using gridded FFTs will not work with this technique, requiring much slower 'direct Fourier transforms'.

4.9. Calculation errors

Computational errors take many forms. Errors in baseline coordinates cause spatially dependent errors in the image. Gridding errors, due to the coarseness of the (u, v) grid, have a similar effect, although they are reduced if each (u, v) cell is well filled. Aliasing of sidelobes and of sources outside the image, due to the resampling operation which enables use of the Fast Fourier Transform algorithm, is often important, especially in smaller databases which have vastly less data than the number of (u, v) cells to be filled. The use of non-gridded subtraction, as in the AIPS task 'IMAGR', can eliminate, or at least greatly reduce, these forms of errors. Use of 16-bit integers in imaging gives an interesting, and unnecessary, limitation of approximately 65,000 : 1 in dynamic range. Tests using 'IMAGR' for perfect data without phase-shifts show that the limit set by computational errors is near $10^9 : 1$.

4.10. Coverage errors

Inadequate (u, v) coverage constitutes an error just as real as any other, and one which is responsible for numerous incorrect interferometric images. Recall that the interferometer is a spatial filter, so that if one is observing a large object with any given array configuration, all information about large-scale structure (approximately larger than 30 synthesized beams) is absent.[2] What one recovers is a spatially filtered image—and often a bad one at that, since the transfer function rarely is smooth. Typically, the shortest spacings 'stick into' the central hole in the (u, v) plane, producing large-scale undulations in the image. If the size of the hole (in wavelengths) is significantly larger than the reciprocal of the source angular size (in radians), then the deconvolution algorithms cannot possibly reconstruct correctly.[3] The result is an incorrect image, with large-scale undulations to boot. In this situation, the only proper solution is to obtain short-spacing data, through observations with a more compact configuration, from another array with the required spacings, or from a single antenna.

5. Techniques of Error Correction

It is clear that antenna- and correlator-based errors are the most important limitations to high dynamic range imaging. I now discuss techniques developed over the past few years that can be employed to accurately remove antenna-based errors and time- and source-invariant correlator-based errors. I will illustrate my remarks with examples of the improvement of images of the well-known quasar 3C 273.

[2] This ratio is appropriate for the VLA. In general, the value is approximately the ratio of the longest to the shortest spacing present in the configuration.

[3] Another way to consider this is to note that important coverage errors will occur when the hole in the coverage is the location of a significant change in visibility. Changes in holes located away from the center of the (u, v) plane may not be critical, since information on the relevant spatial structure is found in other regions. This is not true of the central hole, so that the information in this hole is, in essence, unique.

5.1. Initial editing and calibration

Initial calibration is nearly always performed using observations of nearby unresolved sources. It is obviously advantageous to perform these steps of initial editing and calibration carefully, in order to avoid subsequent problems in imaging. The question of how carefully one should edit is a difficult one to answer in detail. Due to the great robustness of self-calibration algorithms, it is unnecessary to delete any data whose errors are simple multiplicative ones (i.e., involving an antenna phase-shift or gain error). If the source being imaged is weak, so that self-calibration is unlikely to succeed, such points should be deleted. Data involving loss of sensitivity should be deleted if the loss is appreciable. Such an error occurs if the antenna is significantly mispointed, as occasionally happens at the beginning of a scan. Effective procedures for identifying discrepant data include displaying the data in a baseline-time plot, or computing the mean and r.m.s. of each correlator for each scan. Another useful technique is to plot the 'one-dimensional' visibility function—plotting visibility amplitude against (u, v) distance.

Careful perusal of the data at this stage in processing nearly always pays off in quick and efficient imaging. However, it is inevitable that despite the best calibration efforts, important residual errors, especially phase errors caused by atmospheric turbulence, will remain. The only effective procedure to correct these is to employ self-calibration.

Baseline-dependent errors large enough to degrade dynamic range should be flagged if they are time-variable. These errors can be roughly estimated by examining the residuals of the antenna-based gain solution. Keep in mind that background sources will cause apparent closure errors varying from greater than 10% at 90 cm to less than 1% at 2 cm. These will not degrade dynamic range, since they will be correctly handled in imaging. One should be on guard for large, sporadic errors, and delete the appropriate data. Non-variant, baseline-dependent errors can be handled by the techniques of Section 5.3. Before flagging in this manner, be sure to calculate, using the concepts presented in Section 3, the magnitude of the error which will be important. If the expected dynamic range is low (say, $< 100:1$), then the tolerance to a closure error is very high, and correlator flagging will usually be unnecessary, and often undesirable.

5.2. Antenna-based error correction using self-calibration

If the noise on an initial image is significantly above that expected, and if there is sufficient signal, self-calibration is feasible and useful. As stated in Lecture 10, fast convergence of self-calibration is a function of the correctness of the input model. This, in most cases, takes the form of a set of 'CLEAN' components derived from the initial image. However, this is not necessarily the best model. In many cases, the object is dominated by an unresolved core, such that the longer spacings resolve out any associated emission to high degree. In other cases, a model image taken from other data at the same frequency but at different resolution will suffice. Occasionally, an image from a different frequency can be used. The point is that an external image, or model, if available, is often the best way to start things off. Examination of the visibility plot can be extremely helpful in setting an initial model. A good example is provided by 3C 273, as

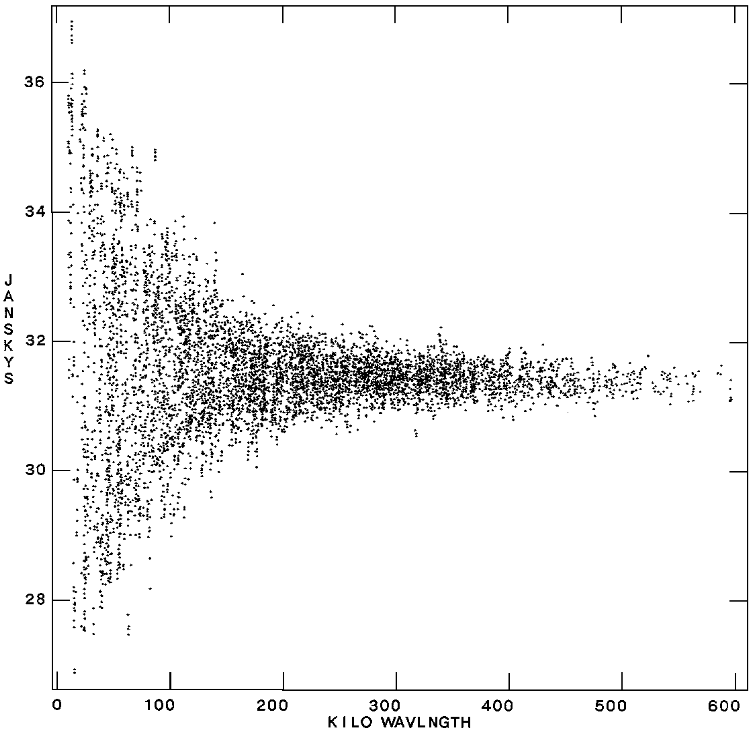


Figure 13–1. The visibility plot of 3C 273 at 6 cm in the **A** configuration. The 'trumpet horn' shape is indicative of a point-dominated source with larger-scale structure that is completely resolved out on longer spacings. A point source of flux density 31.5 Jy provides an excellent initial model for self-calibration.

illustrated in Fig. 5.2.. Due to the large quantity of data, only 10% of the data are actually plotted—this is, however, sufficient to illustrate the main points.

This plot clearly shows the signature of a core-dominated source with associated secondary structure. This structure is essentially totally resolved out for spacings in excess of 200 kλ—the scatter of $\sim$ 5% is due to closure errors. Furthermore, one can see that there are no wildly erroneous data. (Although 90% of the data are not plotted, the important errors are those which repeat, and these will be displayed. Single erroneous values have little effect and can be removed later.) The fact that this source is core-dominated allows an excellent initial model for self-calibration—a point source with 31.5 Jy flux density. This model is nearly perfect, provided that only longer spacings are utilized in the solution. In this case, since the flux density is known, the self-calibration can include both amplitude and phase in the first pass. Note the loss, in this method of self-calibration, of any absolute positional information provided by the original phases.

However, this example is unusual. The more normal situation deals with an extended source, in which case one must make a image, and deconvolve it to provide the model. In this case, the usual, and rather conservative, prescription is to include in the model only those 'CLEAN' components preceding the first negative component. Because this procedure will rarely recover the total flux

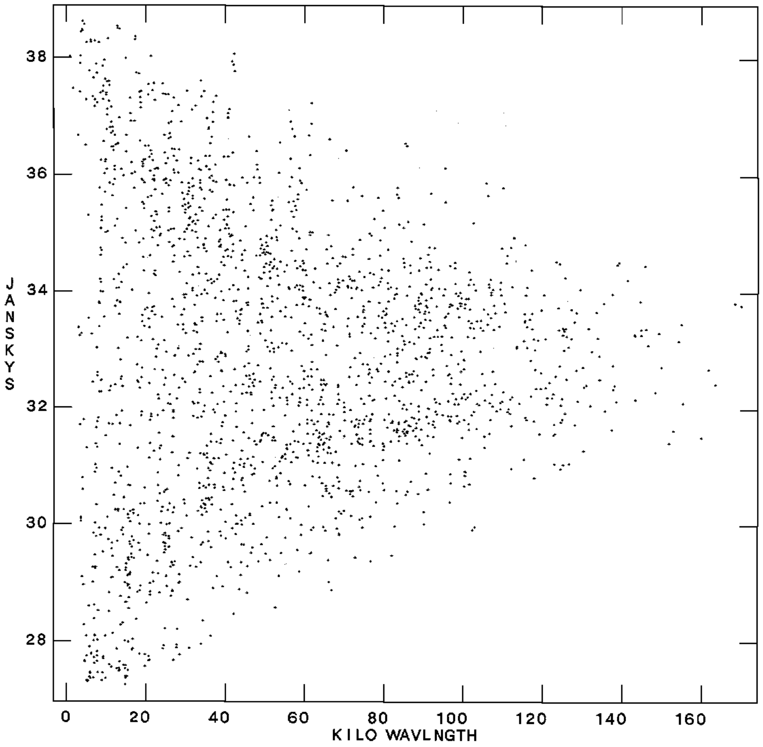


Figure 13–2. The visibility function of 3C 273 at 6 cm in the **B** configuration. Here the secondary noted in the previous figure is not resolved out, so self-calibration will benefit by a model more sophisticated than a point source.

contained in the short spacings, one simultaneously applies a restriction in the (u, v) spacings used, so that only those spacings whose visibility amplitude is less than the total provided by the model will be used. In addition, because poor phase stability can 'lose' flux, the first round of self-cal usually is a phase-only solution. For example, if, in my example, I had chosen to employ a 'CLEAN' model, and the first negative component had been number 11, at which point 33 Jy had been removed, I would have used 10 components as the model, with a (u, v) restriction of 120 kλ to 650 kλ. In this case, the safe approach is to calibrate the phases alone. After this, a new image would be created, and the subsequent deconvolution should produce more positive components than before. This initiates a second round of self-calibration, with both amplitude and phase solutions. When employing amplitude solutions, it is usual and advisable to 'float' the gains—renormalize the gain solutions so that the mean solution is of unit magnitude. This prevents the gain solution from being affected by the model having too little flux, thus systematically decreasing the total apparent flux density of the source.

On occasion, the self-generated initial model is so poor that there is little hope for fast convergence. What then? This situation occurred for the **B** configuration data of 3C 273. The visibility plot for this configuration is shown in Fig. 13–2. A point-source model is obviously a poor choice here, as the secondary is not resolved out on any baseline. The usual procedure is to make a dirty image and deconvolve – however, in this case, the first negative 'CLEAN' component came before the first component from any point other than the core, so the usual prescription would return only a point-source model. This situation was the result of some very poorly edited data—I was was a little lax in my standards, since I was sure that self-calibration would work! Under these circumstances, the **A** configuration image was used to self-calibrate the **B** configuration data, with a (u, v) restriction to baselines longer than 50 kλ.[4] Because the core flux density had changed between epochs of observation, a phase-only solution was made.[5] The resulting improvement was spectacular, as demonstrated in Fig. 13–3. This gave a much improved model for the second round of self-calibration, allowing phase and amplitude gains to be simultaneously calculated. The result of this is also shown in Fig. 13–3. It will be immediately apparent that the noise is not uniform. This is generally true for all objects, but especially so for this object since its declination is 2°. The distribution of noise must follow the beam sidelobes, since the origin of the noise is limited to sampled (u, v) cells. At most declinations, the distribution of filled cells is sufficiently uniform to allow the noise distribution in the image to appear uniform. However, at low declinations, all tracks are nearly E–W, so the resultant noise is primarily distributed N–S. At high declinations, all tracks are circular, and similarly the resulting noise patterns are nearly circular.

Experience shows that after two or three loops of self-calibration, improvement is rather slow. The limitations to dynamic range are now primarily due to baseline-dependent errors. These take two forms. The first are those which are due to a few very bad points—due, for example, to weak interference or correlator malfunctions. The second are what I would term 'true' closure errors, slowly varying, multiplicative in amplitude and additive in phase. The truly discrepant points can be quickly identified by subtracting the (inverse) Fourier transform of the model from the data, thus reducing the database to those residuals which are in strong disagreement with the best image.[6] The largest discrepant values can then be easily identified by plotting the residuals, and removed by flagging. The model can then be put back in (adding the inverse Fourier transform of the model to the data), and a new image made. Figure 13–4 shows an example of this procedure.

Some bad values are clearly present. At this point, I recalled that the antenna-based gain solutions for the calibrator persistently complained about high and variable closure errors for all baselines attached to one antenna/IF. Closer inspection revealed that the data from baselines formed from this antenna/IF were erratically variable. The decision made at that time was to keep the data, in case further processing could allow correction. Since current software cannot handle time-variable baseline-dependent errors, the data from this

[4] This restriction is not required but is a useful precaution, as the **A** configuration undersamples the (u, v) plane in this region.

[5] The visibility phases are less sensitive to a change of flux than are the visibility amplitudes.

[6] Note that the '.IMAGR' file in the AIPS program 'IMAGR' contains these residuals in a form which can efficiently be plotted.

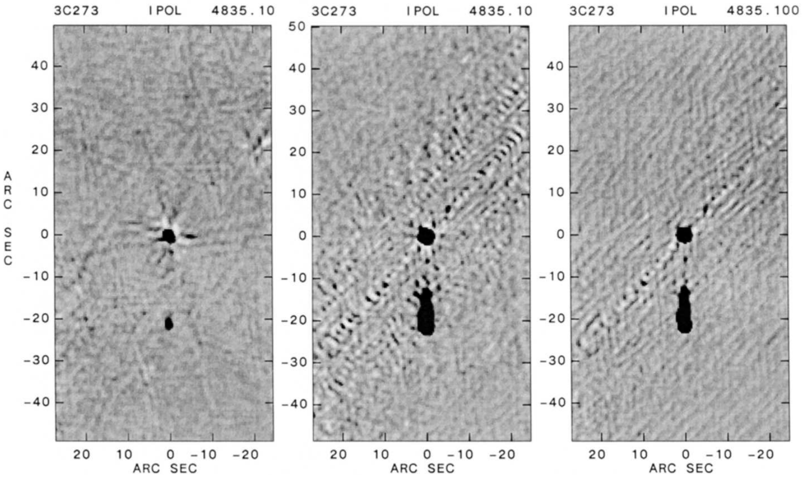


Figure 13–3. Images of 3C 273, made from **B** configuration data, demonstrating three stages of self-calibration. All three images have been rotated to make the structure vertical. *(Left)* The image without any self-calibration. The greyscale extends from −1 to 1 Jy/beam. The peak is 28.2 Jy/beam, and the r.m.s. noise is 134 mJy/beam. *(Center)* The image after self-calibration using the **A** configuration image as an input model, correcting phases only. The greyscale extends from −25 to 25 mJy/beam. The peak is 32.9 Jy/beam, and the r.m.s. noise is 5.5 mJy/beam to the North and South, 2.3 mJy/beam to East and West of the core. *(Right)* The image after a second self-calibration iteration, using the center image as a model, and solving for both amplitude and phase. The greyscale is the same as the center, and the r.m.s. noises are 4.5 and 2.3 mJy/beam in the directions indicated before.

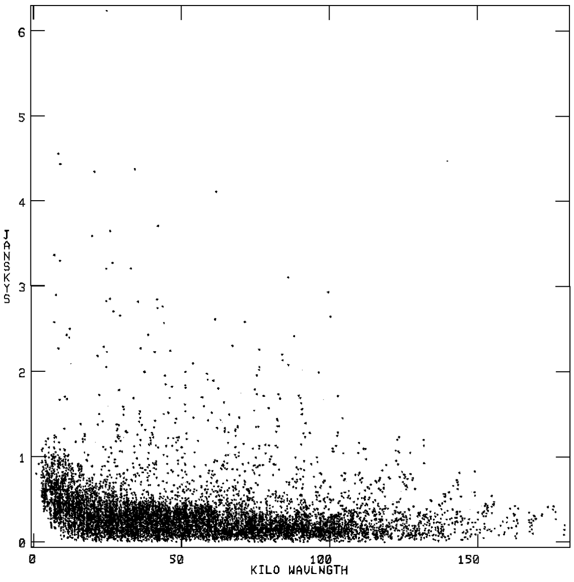


Figure 13–4. A plot of the visibility amplitudes after the inverse Fourier transform of the model has been subtracted. The points lying under 1 Jy represent normal residual closure errors, while the points scattered above this are all due to a malfunctioning correlator.

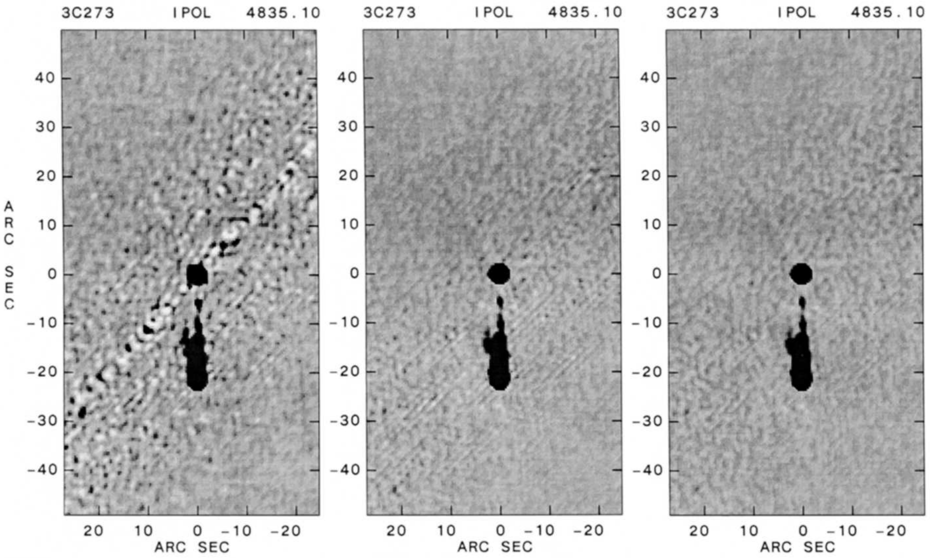


Figure 13–5. Images of 3C 273 after further stages in processing. All are shown with greyscale wedges from −10 to 10 mJy/beam. *(Left)* After self-calibration, followed by removal of a strongly variable correlator. The r.m.s. noises are 2.5 and 0.9 mJy/beam in the N–S and E–W directions respectively. *(Middle)* After application of closure corrections calculated from the source itself. The noises are 1.23 and 0.5 mJy/beam. *(Right)* After clipping residual visibilities, thus removing the largest time-variable closure errors. The noises are 1.03 and 0.42 mJy/beam.

antenna/IF were then flagged. The subsequent image, shown in Fig. 13–5, improved dramatically. Considerable time and effort could have been saved had I done the required flagging initially.

The techniques described above are the main tools for antenna-based self-calibration. Some minor refinements, dependent on user, exist. If you have gotten approximately 40 dB in dynamic range on your image, know that your thermal level lies well below the current noise, and have noticed that repetition of the above steps is achieving little, then you are ready for baseline-based calibration.

5.3. Baseline-based error correction

The principles of baseline-based calibration are identical to those of antenna-based calibration. By observing strong, isolated sources, estimates of the (complex) multiplicative corrections can be made and applied to the data. Since these corrections are much smaller than the antenna calibration corrections (generally less than 1% and 1° for the VLA), they must be done after the best antenna calibration has been completed.

I re-emphasize that the majority of observations will not require correlator calibration. If the limiting dynamic range, set by thermal noise, is less than 1000 : 1 to 10,000 : 1 (depending on the quantity of data), then this calibration will not be effective. Another way of approaching the question is to look at the noise ‘footprint’ on the best image obtained without baseline-based calibration. If traces of the beam sidelobes are still present, this form of calibration is likely to be effective. For example, the residual sidelobes in Fig. 13–3 (right) clearly show the N–S disturbance expected for a low-declination source. These are likely due to persistent correlator offsets. So, before embarking on this form of calibration, be sure the images show the effects expected of baseline-based errors.

Recent tests performed with VLA data from observations of strong calibrators show that the correlator errors are present at levels of approximately 0.5% in amplitude, and $0\overset{\circ}{.}5$ in phase. The distribution of errors is non-Gaussian, with a few baselines showing errors of 3 to 5 percent/degree. Following the arguments of Section 2.2, it is reasonable to expect that these errors are responsible for the low apparent dynamic range. Furthermore, these errors are slowly time-variable. The software available at the present time allows only time-invariant solutions—however, the results are generally encouraging, so it appears that the time-variable part is less important than the mean level, on the time scales of interest.

The procedure for calibration of these errors parallels that for the antenna-based calibration. Observers wishing to include closure correction must observe a very strong calibrator. Closure correction calculations are almost always noise-limited—simply because the desired correction is of order 0.1% of the amplitude, and must be done baseline-by-baseline. Simple application of the radiometer equation given in Section 2.3 shows that calibrator flux density is of the utmost importance. Use of phase calibrators (typically of 1 Jy flux density) is not as effective as two or three observations of 3C 286 or 3C 48. An important point is that ALL the structure of the calibrator, including all background sources, must be included in the closure calculation. Typically, the effect of background sources is similar to that of the closure errors. Structure not included in the process will show up on the resulting image. Within AIPS, the task ’BLCAL’ can be used to determine the baseline-dependent errors. See the AIPS Cookbook for guidance on how to use ’BLCAL’.
Some results of applying this procedure to observations of 3C 273 are shown in Figures 13–5 and 13–6. In almost all situations one must use a strong, isolated, simple source for these corrections. However, for this case, 3C 273 itself satisfies these criteria, so I have used it to calculate its own closure errors (a procedure called by some ‘incestuous self-calibration’). The reason for using a simple isolated source for closure corrections is to prevent ‘pre-defining’ the structure.[7] This danger is reduced by restricting the solutions to be time-averaged. However, inadequacies in the model show up as excess flux in the short spacings, which also have the slowest fringe rates. Time-averaging the residuals will be effective only if the averaging interval is many times the fringe period. In **D** configuration this resulting time span will often be impossibly long. Thus, use of the source itself for closure corrections is dangerous, and should always be avoided. The only exception is when the model clearly includes all the flux

[7] To better grasp this problem, note that closure corrections are modifying every visibility. If a time-varying correction is calculated from a model of a source, and applied back to the data at each time the correction is calculated, then clearly the data will be modified to exactly reproduce the model. Thus, for example, a double source could be turned into an unresolved point source.

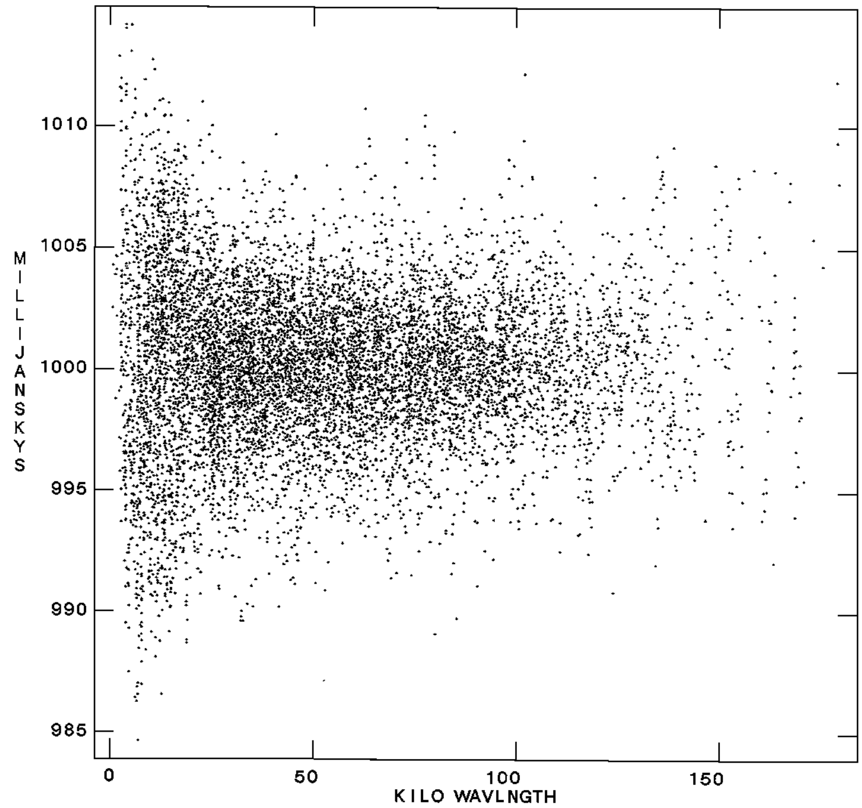


Figure 13–6. The visibility after division by a model provided by the image after self-calibration. The scatter about 1.0 is due to multiplicative closure errors.

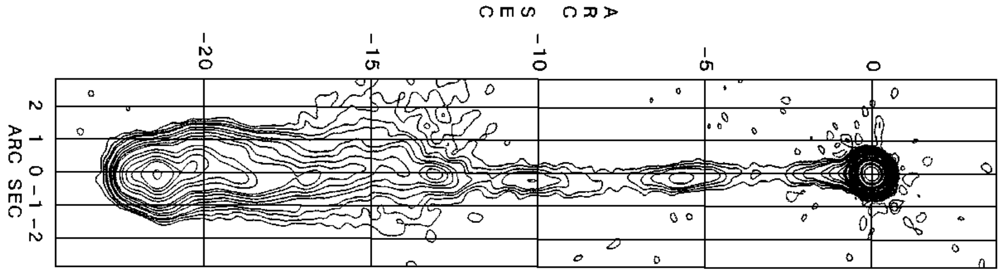


Figure 13–7. 3C 273 at 6 cm observed with the **A** configuration and the spectral line correlator. The only required processing was a single self-calibration, and a single closure correction—which showed the closure errors to be less than 0.1%. The dynamic range is better than 200,000 : 1.

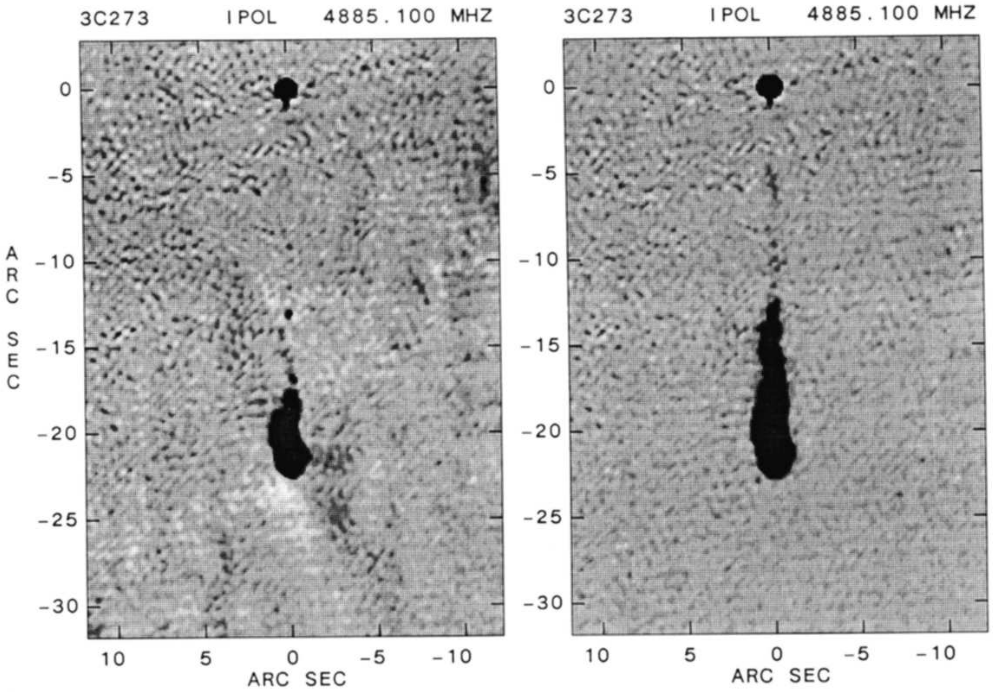


Figure 13–8. An illustration of the effect of missing (u, v) coverage on a synthesis image of 3C 273 at 6 cm. In both panels the greyscale runs from -10 to 10 mJy/beam. The left panel shows the image using **A** configuration data only, so that structures of scale $\geq 7''$ are severely attenuated. Note the large-scale undulations in the noise around the source. The right panel shows the image after adding the **B** configuration data. The background undulations are completely absent, and much missing source structure has appeared.

density. The result of dividing the data by the inverse Fourier transform of the model, as shown in Fig. 13–5(left), is shown in Fig. 13–6. Any unmodeled flux shows up as a drift of the mean ratio away from unity—so in this case it appears that all the flux is represented. Note that the closure levels are as expected, generally less than 1%. Application of the mean correlator-based offsets to the data results in the image shown in the center panel of Fig. 13–5. Especially note the almost complete disappearance of the N–S disturbance—good evidence that the closure errors have been greatly reduced. The final step performed was a second subtraction of the inverse Fourier transform of the model from the data, followed by deletion of the largest remaining residuals. Restoration of the model results in the image shown in the right-hand panel of Fig. 13–5. The dynamic range here is 78,000 : 1.

The dynamic range of the image shown in Fig. 13–5 is good, but still at least an order-of-magnitude from theoretical. What next? The suspicion is that small, time-variable baseline-dependent errors are the next limiting factor. As I have already stated in Section 4.3, the problem of 'Closure Errors' can be avoided by using the VLA's spectral line correlator. Tests done with this have given dynamic ranges exceeding 150,000 : 1, with considerably less effort involved than that used to get a continuum correlator image up to 78,000 : 1. Figure 13–7 shows the result of such an observation of 3C 273 at 6-cm in the **A** configuration. The dynamic range is approximately 215,000 : 1. Clearly, use of this mode is to be preferred for high dynamic range imaging. However, users should be aware that a considerable loss of bandwidth results, so the *potential* dynamic range is notably less than is possible with the continuum mode.

5.4. Coverage errors

It might be thought that the solution to this problem is trivial, and indeed it often is, if getting and calibrating more data is to be considered such. However, there are some subtleties, which I will briefly comment on here.

I first demonstrate how this 'error' can affect you. In Fig. 13–8 is shown an image of 3C 273, taken in the **A** configuration after the very best self-calibration and closure corrections. Notice that the noise is not 'flat', but that there are large-scale undulations radiating away from the extended component. These are a result of inadequate short-spacing (u, v) coverage. Recall that any given VLA configuration covers adequately a range of about 20 in resolution. That is, any structure larger than about 20 times the angular extent of the synthesized beam will be significantly attenuated, with accompanying errors. For this example, the resolution is approximately $0\rlap{.}''35$, so any structures larger than approximately $7''$ are suspect. This is the scale of the secondary. The solution is simple: get some **B** configuration data.

In principle, combining the data from two configurations is simple, requiring only that the basic calibration be correct, so that the two databases have the same amplitude scale. Different phase centers can be handled, provided that the shift is a small fraction of the primary beam size (so that objects near the image edge do not appear time-variable). However, there is one complication. Core-dominated sources, such as 3C 273, are time-variable, so that the source structure actually changes in time (violating the first principle of aperture synthesis). Fortunately, the solution is simple if the changes in flux density are known: one can merely subtract the difference from one database. In practice, the position of the variable component must be accurately known, so individual self-calibration of the databases is required. This requirement means that the subtraction should be done on the data from the higher resolution configuration. Subsequent concatenation of the two databases should not be done unless they are in the same phase frame. The easiest way to guarantee this is to perform 'cross- self-calibration', using the model from one configuration to self-calibrate the other. This actually works for adjacent configurations of VLA data! Another approach, often better, is to combine the databases, make an image, and use this for self-calibration of each database.

When combining data taken from different arrays, a common question is whether one should use the high-resolution data to self-calibrate the low-, or vice-versa. The answer depends upon circumstances. For core-dominated situations, it is clearly advantageous to start with the high-resolution data, since in this case a simple and accurate model for the data is readily available. However, for large, complicated objects such as Cygnus A, I have found the reverse procedure to be much more effective. The **D** configuration data in this case were of excellent quality, and they provided an excellent model for the **C** configuration data, and so on. Again, the best rule-of-thumb is to start with the best model available.

Another, related question is one of *bandwidth synthesis*. This is the technique of observing at frequencies that differ (by, say, 10%) to improve the (u, v) coverage. It can be highly effective for large objects when the (u, v) plane is inadequately sampled at one frequency alone. However, this technique may be dangerous for sources with large spectral index gradients. This topic is the subject of Lecture 21.

References

Bagri, D. S. 1990, VLA Electronics Memorandum No. 216, NRAO.

Clark, B. G. 1978, VLA Electronics Memorandum No. 171, NRAO.

D'Addario, L. R., Thompson, A. R., Schwab, F. R., & Granlund, J. 1984, *Radio Science*, 19, 931–945.

Schwab, F. R. 1979, VLA Computer Memorandum No. 150, NRAO.

Thompson, A. R. 1980, VLA Electronics Memorandum No. 192, NRAO.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

14. Image Analysis

Ed B. Fomalont
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A.

Abstract.

Image analysis is the general term applying to those procedures and techniques which are used to interpret and parametrize information from an image or a set of images. These procedures also include obtaining error estimates for the derived parameters and estimating the image reliability. Image analysis is a vague term, and the choice of image analysis techniques is dependent on the nature of the observations and the type of questions which motivated the experiment. Nevertheless, some general analysis techniques are useful to discuss. The philosophies of most of the techniques will be emphasized, and implementation details will not be discussed except when necessary. VLA software will be mentioned in connection with specific algorithms.

Here it is assumed that the set of images has been appropriately processed and is of good quality. For aperture synthesis, this processing includes data editing and calibration (Lecture 5), as well as deconvolution of the point spread function (Lecture 8) and self-calibration (Lecture 10), if necessary. Apart from those defects which are peculiar to aperture synthesis, much of the material of this lecture should be applicable to images from a variety of astronomical instruments.

Image display is an important aid in image analysis. For simple images, grey-scale and color TV-oriented displays and contour diagrams are essential for determining the general features in the intensity distribution which are amenable to analysis. For complicated images, particularly sets of images over frequency, subtle and ingenious displays are required to perceive faint features and morphologies. Once recognized, these features can be analyzed and parametrized in a manner which is astronomically useful.

Whenever a specific image analysis function is described in this lecture, the name of a computer implementation of that function, within the NRAO Astronomical Image Processing System (AIPS), will be mentioned also.

1. Image Modification

Several types of image modification are useful in analysis and display. Two that are described in this section are: convolution, which smooths or sharpens an image; and interpolation, which modifies the grid on which the intensities are defined. Other general types of image correction are also mentioned.

1.1. Smoothing or sharpening an image

The derived intensity distribution, after appropriate data reduction and imaging techniques have been applied, represents an estimate of the true intensity distribution in the region of sky that is of interest. Because of the finite resolution of the measuring instrument, the derived image is a smoothed version of the true image (distortions and noise also remain in the derived image). The resolution of the image can be further modified in order to better discern very small or very large features. Figure 14–1 shows a contour display of a radio image which contains large-scale and fine-scale features. It is obvious that different image properties are better suited for measurement at different resolutions. The overall appearance of the complex features is seen in image (a), the integrated intensity is most reliably calculated from image (b), and the parameters associated with the bright feature are best determined from image (c).

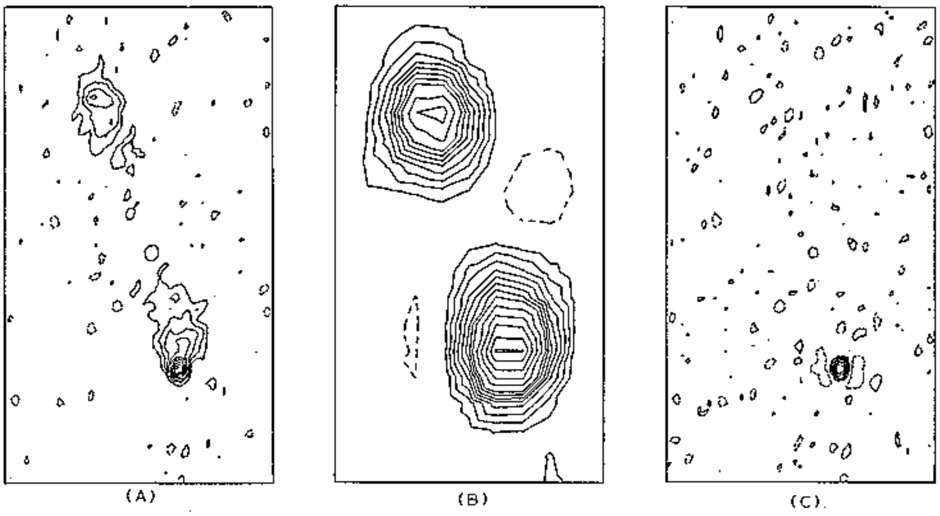


Figure 14–1. Relative emphases of image features achieved through use of differing resolutions. **(a)** 0.″5 × 0.″4 resolution; **(b)** 2.″0 × 2.″0 resolution; **(c)** 0.″5 × 0.″4 resolution, with high-pass filtering.

There are several methods that may be used to modify the resolution of an image. The most straightforward method is to convolve the image $I(l)$ with an appropriate kernel function $K(l)$, to obtain the modified image $I'(l)$. In nearly all cases the intensity is defined on a regularly spaced grid, with abscissae l_i, and in one dimension the convolution is

$$I'(l_j) = \frac{1}{N} \sum_i I(l_i) K(l_j - l_i) \,, \tag{14–1}$$

where N is a normalization factor. Extensions to n dimensions are obvious. Some examples of convolution functions are

	$K(\|l_i - l_j\|)$		
$\|i - j\|$	(1)	(2)	(3)
0	1.0	1.0	1.0
1	0.5	0.9	−0.4
2	0.0	0.5	0.3
3	0.0	0.3	−0.2
4	0.0	0.1	0.1 .

The use of kernel (1) produces a slight convolution called Hanning smoothing; kernel (2), a heavier smoothing; and kernel (3) sharpens small-scale features. Each kernel is symmetric about the origin; Gaussian-shaped kernels are commonly used. The normalization factor N depends on the nature of the convolution. If $N = 1$, the intensity scale of the image is unchanged. If N equals the integral of the kernel, then the integrated intensity is unchanged.

A related method of convolution uses Equation 14–1 transformed into spatial frequency space; I shall use u to denote a point in this space. If $V'(u)$,

$V(u)$, and $k(u)$ are the (inverse) Fourier transforms of $I'(l)$, $I(l)$, and $K(l)$, respectively, then the convolution formula becomes

$$V'(u_j) = V(u_j)k(u_j)\,. \tag{14–2}$$

The function $k(u)$ can be interpreted as a weighting factor in spatial frequency space. The functional form of a kernel is often that of a Gaussian function, since its Fourier transform is also Gaussian. This method is easily applicable to those instruments where the (inverse) Fourier transform of the image is directly measured (see use of the parameter 'UVTAPER' in AIPS task 'IMAGR'). Even with the extra overhead of two Fourier transforms, many convolutions are more quickly calculated using Equation 14–2. For example, a high-pass filter is obtained by setting $k(u_j) = 0$ for $u_j < U$, and $k(u_j) = 1$ for $u_j > U$. The associated kernel $K(l_i)$ is an oscillating function similar to kernel (3), except that many terms must be used and the computation time is therefore much longer than with the Fourier reweighting method.

Another convolution method is that associated with the deconvolution/reconvolution technique employed by the AIPS tasks 'APCLN', 'IMAGR' and 'VTESS' (see Lecture 8). The deconvolution part of these tasks decomposes the observed image (often called the 'dirty' image) into a set of point components which, when convolved with the point-spread function of the observations (the 'beam' pattern), reproduce the observed image. This set of point components can then be reconvolved with any desired beam pattern to produce an image (often called the 'CLEAN' image) which does not contain artifacts that were present in the original beam. Generally, a truncated Gaussian-shaped beam is used. In principle, the point components can be convolved to any desired resolution, including *super-resolution*.

However, the 'CLEAN' image is not a precise convolution of the true intensity distribution. The part of the image which has been decomposed into a collection of point components has a different resolution than that of the residual image (which includes both residual sky signal and noise) which was not decomposed in the 'CLEAN' algorithm. The difference is minimized in two ways: First, the reconvolution function used for the point components should be similar in shape to the original beam pattern. Second, the decomposition should proceed to low levels in the dirty image, well down into the noise level, so that the low-level intensity and the larger noise contributions are included in the point components and then reconvolved to the 'CLEAN' beam.

For strong, isolated features the use of a radically different reconvolution function from the original beam pattern is an inexpensive method of smoothing the feature. It is possible to obtain images with keener resolution than the instrumental limit by convolving the point components with a kernel which is narrower than the original beam. The results can be misleading and are accurate only for bright, isolated features.

1.2. Interpolating an image

The image intensity distribution is generally defined over a rectangular lattice specified at an early stage of reduction. The calculation of the image intensity at an arbitrary point or on a new grid of points is necessary for a host of image analyses and displays. Several obvious applications are: determination of the position of isolated features; alignment of a set of images onto the same grid; and mosaicing a set of small conterminous images into one large image.

If $I(l_i)$ represents an intensity distribution defined on a grid, then the interpolated intensity $I'(l')$, where l' is an arbitrary point, is also given by Equation 14–1, with l_j replaced by l'. The argument of the kernel is not, in general, an integer. If the intensity distribution is band-limited—i.e., contains no frequencies higher than U—then a perfect interpolation kernel is $K(z) = \frac{\sin 2\pi Uz}{2\pi Uz}$. For an image which is sparsely sampled, Δl is approximately $\frac{1}{2U}$, and interpolation with this kernel over a large domain is expensive to calculate. If the image is well-sampled, $\Delta l \ll \frac{1}{2U}$, then the adjacent points are not independent, and a simpler, smaller kernel will produce an accurate interpolation (AIPS tasks 'GEOM' and 'HGEOM'). Some examples are truncated sinc functions, splines, the Everett linear function, etc. (Weast & Selby 1975). There is, however, a slight change of resolution with some of the interpolation functions.

If the Fourier transform $V(u)$ of the image has been directly measured, then a regridded image can be made via a new Fourier transform. However, all subsequent processing—from the dirty image to the final image—must be redone.

1.3. Primary beam correction

The antennas which comprise an array are sensitive to radiation coming from a small region of sky. Correction for the relative sky sensitivity across the image (the primary beam correction) is made only after the best-quality image has been obtained. Correction of the image at an earlier stage of reduction (e.g., the dirty image) leads to incorrect results. If the image contains only a few bright features, the correction need be applied only locally.

There are several second-order problems associated with the primary beam correction at the VLA. First, for antennas with azimuth–elevation mounting, the relative orientation of the sky and the primary beam changes over a period of hours. If the primary beam is not circularly symmetric, then there is no unique primary beam associated with the set of observations. In severe cases (for very extended sources and large primary beam ellipticity), images made over long periods of time must be separately processed over periods that are short compared with changes in the sky–primary beam orientation. Sometimes an image is contaminated by a strong source outside of the main lobe of the primary beam pattern. Over a period of hours, this source will appear to vary because of the changing antenna sensitivity at the location of the source. It is difficult to remove the artifacts caused by this variability.

Second, the primary beam correction for the four correlation channels (RR, RL, LR, and LL) may differ, as is the case for the VLA. There is a slight offset in the pointing axis between the right- and left-hand circular polarizations, and the Stokes Q and U sensitivities are significantly different from that of the total intensity. If such differences are important, then each correlation must be processed independently and then combined only after correction with its proper primary beam.

1.4. Other image defects

Many second-order image corrections are discussed in Lectures 17, 18 and 19 – for bandwidth smearing, integration-time smearing, non-coplanar baselines, and grid curvature. It is generally very expensive to correct an image for these defects. However, the parameters of discrete features can be corrected in order to compensate for these defects. An example will be discussed in Section 2.1.

2. Parameter Estimation of Discrete Components

Images often contain bright, isolated features—or 'components'—whose essential characteristics can be represented by a few well-defined parameters. Accurate error estimates can often be derived for these parameters, and the features then can be easily compared at different frequencies and with data taken at different epochs. The simplest set of parameters defining the component properties consists of the moments of the distribution (AIPS tasks 'MOMNT' and 'MAXFIT'). In one dimension the (first three) moments are

$$\begin{aligned} F &= \sum I(l_i) && \text{Integrated Intensity,} \\ L &= \tfrac{1}{F}\sum l_i I(l_i) && \text{Mean Position,} \\ B &= \sqrt{\tfrac{1}{F}\sum (l_i - L)^2 I(l_i)} && \text{Width;} \end{aligned} \qquad (14\text{–}3)$$

here they have been normalized in the usual manner. A weight associated with each pixel can also be incorporated in the moment calculations. Extension to two dimensions is straightforward. Higher-order moments can be defined, but they are of little use for most astronomical applications.

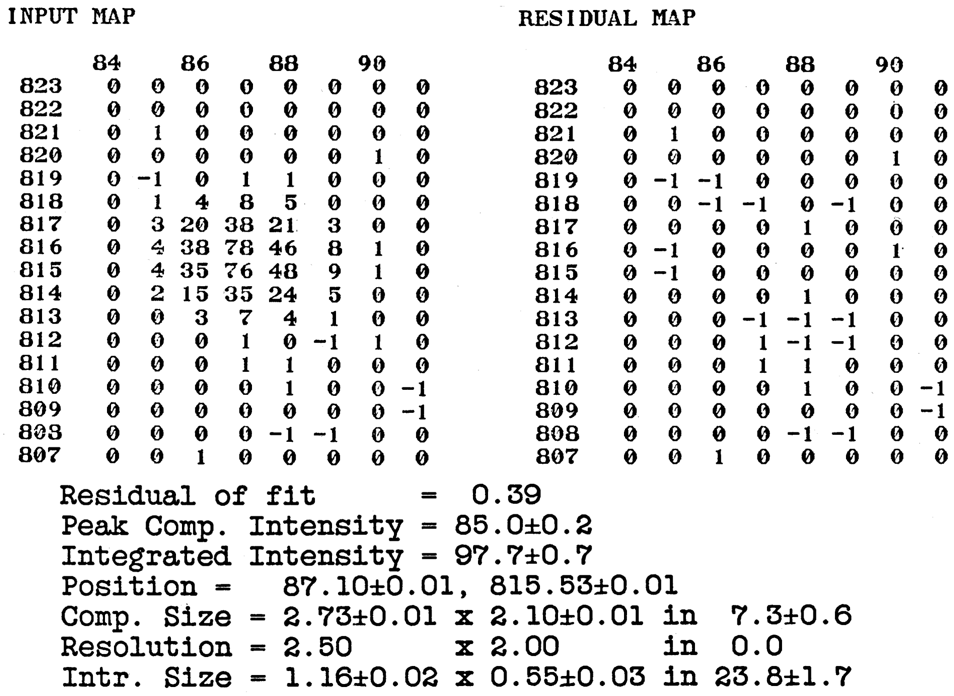
2.1. Model-fitting

It is often more convenient to determine parameters for a component of an image within the framework of a specific model intensity distribution. For components which are not heavily resolved, the point-spread function is generally chosen as the model, and slightly resolved features are decomposed into several displaced model components. For a dirty image, the point-spread function is the dirty beam; for a 'CLEAN' image, the point-spread function is approximately Gaussian-shaped. An image of an extended object of known shape (such as a planet) can be fit to a model consisting of a uniformly-illuminated circular disk described by several appropriate free parameters.

After selection of the appropriate model M for the component intensity distribution, with free parameters p_j, the model intensity distribution $M(p_j;i)$ is calculated at the grid points l_i around the component. The parameters and error estimates are determined by the method of maximum-likelihood. If one assumes that the errors are distributed normally, then the method is equivalent to minimizing the variance, σ^2,

$$\sigma^2 = \sum (M(p_j;i) - I(l_i))^2 \,. \qquad (14\text{–}4)$$

Many fitting techniques are available, and these depend on the analytical tractability of the model functions. Since the free parameters are not generally orthogonal, even in the one-dimensional case, nonlinear fitting methods must be



```
INPUT MAP                                         RESIDUAL MAP

      84      86      88      90                        84      86      88      90
823   0   0   0   0   0   0   0   0             823   0   0   0   0   0   0   0   0
822   0   0   0   0   0   0   0   0             822   0   0   0   0   0   0   0   0
821   0   1   0   0   0   0   0   0             821   0   1   0   0   0   0   0   0
820   0   0   0   0   0   0   1   0             820   0   0   0   0   0   0   1   0
819   0  -1   0   1   1   0   0   0             819   0  -1  -1   0   0   0   0   0
818   0   1   4   8   5   0   0   0             818   0   0  -1  -1   0  -1   0   0
817   0   3  20  38  21   3   0   0             817   0   0   0   0   1   0   0   0
816   0   4  38  78  46   8   1   0             816   0  -1   0   0   0   0   1   0
815   0   4  35  76  48   9   1   0             815   0  -1   0   0   0   0   0   0
814   0   2  15  35  24   5   0   0             814   0   0   0   0   1   0   0   0
813   0   0   3   7   4   1   0   0             813   0   0   0  -1  -1  -1   0   0
812   0   0   0   1   0  -1   1   0             812   0   0   0   1  -1  -1   0   0
811   0   0   0   1   1   0   0   0             811   0   0   0   1   1   0   0   0
810   0   0   0   0   1   0   0  -1             810   0   0   0   0   1   0   0  -1
809   0   0   0   0   0   0   0  -1             809   0   0   0   0   0   0   0  -1
803   0   0   0   0  -1  -1   0   0             808   0   0   0   0  -1  -1   0   0
807   0   0   1   0   0   0   0   0             807   0   0   1   0   0   0   0   0
     Residual of fit        =  0.39
     Peak Comp. Intensity = 85.0±0.2
     Integrated Intensity = 97.7±0.7
     Position =    87.10±0.01, 815.53±0.01
     Comp. Size = 2.73±0.01 x 2.10±0.01 in  7.3±0.6
     Resolution = 2.50      x 2.00      in  0.0
     Intr. Size = 1.16±0.02 x 0.55±0.03 in 23.8±1.7
```

Figure 14–2. Image fit to a bright component.

used. Most methods converge more quickly, and to a reasonable solution, if an accurate initial guess of the model parameters is supplied. Additional parameters, such as a zero-offset in the fitted region, may be included to incorporate image defects which are not associated with the component.

An example of a fit of a strong feature to a Gaussian-shaped component is shown in Figure 14–2 (AIPS tasks 'IMFIT' and 'JMFIT'). The feature is only slightly resolved. The residual image (image minus model) suggests that the fit is reasonable; the scatter of points near the component is about the same as that over the entire region. For features whose peak intensity is less than about five times the r.m.s. fluctuations, some fitting algorithms may produce biased parameter estimates.

Fitting multiple components to a complex feature is often a waste of time. The shapes of the sub-components are not generally known (why should they be Gaussians, for example) and the parameter values obtained will be tightly coupled. A practical limit in the complexity of a feature for reasonable model-fitting is two overlapping, not too extended, components with perhaps a point component somewhere in the feature. This model, using two-dimensional Gaussian components, contains 13 components. It may still be necessary to hold some of the free parameters constant for reasonable convergence.

2.2. Parameter adjustments

Once a satisfactory fit is obtained, parameters may be corrected for known image properties or defects. The most common correction is that for the finite resolution of the image. The true width of the component can be obtained using Equation 14–4, by taking the difference between the second moment of the fit and that of the point-spread function. For a 'CLEAN' image with a Gaussian-shaped beam and model distribution, the true component size can be determined analytically. As illustrated in Figure 14–2, removal of the image resolution from the fitted width gives an estimate of the size of the component with the primary instrumental resolution removed.

An example of correcting for another instrumental effect is one associated with bandwidth smearing (see Lecture 18). The component which was fit in Figure 14–2 is displaced 70″ in position angle 30° from the phase center of the observations. With a 25 MHz bandwidth, the expected bandwidth smearing is 1.2 pixels in the direction of the displacement (see Lecture 18). Thus, the real width of the component along the direction of displacement from the phase center is about that expected from bandwidth smearing; the component is probably unresolved (< 0.2 pixels). The combination of the radial (bandwidth) smearing and the Gaussian beam produces an image shape that is non-Gaussian. More exact methods of analysis are possible but not always necessary.

Adjustment of parameter amplitudes is common. Corrections of component widths caused by instrumental effects do not, to first order, change the integrated intensity in the component, and correction of the peak intensity of the component must be made. Another example is the correction for primary beam attenuation, which was discussed in Section 1.3.

2.3. Parameter errors

Error estimates obtained directly from fits should be viewed with skepticism. There is generally a built-in assumption that the image errors are stochastic and independent over the component, which may not be valid for a variety of reasons. Let ΔR be the post-fit r.m.s. error associated with the pixels in the image. Then the approximate errors of the parameters (Z = Zero Level, P = Peak Intensity, L = Position, W = Width, F = Integrated Intensity) for one component in one dimension are:

$$\begin{array}{rcl} \Delta Z = & \Delta R/3 & \text{Zero Bias,} \\ \Delta P = & \Delta R & \text{Peak Intensity,} \\ \Delta L = & \Delta R W/2P & \text{Position,} \\ \Delta W = & \Delta R W/P & \text{Width,} \\ \Delta F = & \sqrt{\Delta R^2 + (I \Delta W/W)^2} & \text{Integrated Intensity.} \end{array} \qquad (14\text{–}5)$$

The above expressions are rough guides, and the true errors may be larger. When fitting in two dimensions, similar expressions apply. There is often a strong correlation between the derived integrated intensity and component diameters. When fitting several blended components, many parameters are not well separated, and their associated errors become much larger than the above guidelines would indicate.

2.4. Fitting models to the visibility data

If the image quality is poor, it is sometimes preferable to compare the visibility data directly with the (inverse) Fourier transform of a well-defined, *a priori* model. Reasons that might account for poor image quality are:

1. the paucity of input data, causing the synthesized beam to be of poor quality,
2. the inaccuracy of the measured visibility phase, and
3. the distortions of very large features in the image.

Circumstances under which this type of fit are useful include: the determination of the positions of strong isolated components using only the visibility phases; fitting the visibility amplitude data of a planet or star (i.e., speckle data) to a disk model or the outer portion of the Sun to a limb model; and determining the size of small features with few visibility data samples (AIPS task 'UVMOD'). However, these model-fitting techniques are not useful for data of low signal-to-noise ratio, and the ambiguity of the fit to complicated models is often a problem.

The claim is often made that it is possible to obtain higher resolution on strong sources by applying model-fitting techniques directly to the visibility data, rather than by applying analysis methods to the associated image, even when the three limitations discussed above do not apply. Two examples are determining accurate positions of point-like objects and determining the angular size of barely resolved sources. However, analysis of a properly made 'CLEAN' image via good analysis techniques should produce parameter values and errors which are equal to those of model-fitting the visibility data—and with less ambiguity about whether the model is in fact appropriate. The analysis may require emphasis of the outer spacings of the visibility data before imaging, to weight to data in a similar manner as the model-fitting technique. The technique of super-resolution (using a 'CLEAN' beam which is narrower than the dirty beam) may also be useful.

3. Parameter Estimation for Extended Sources

3.1. General problem

Parameters describing extended features are difficult to obtain and are ambiguous to define. Extended features often contain sub-components of various sizes and shapes, and there are often long, thin, curved features. Attempts to fit such a complicated distribution with a myriad of Gaussian components are a waste of computing resources. Fitting of the brighter sub-components does make sense, and here the discussion of the preceding section is relevant. Intelligent image display is needed at this stage to determine which aspects of the image or set of images to analyze. Of course, there are many morphological properties of some images which cannot be parametrized, and a suitable display is all that is needed in these cases.

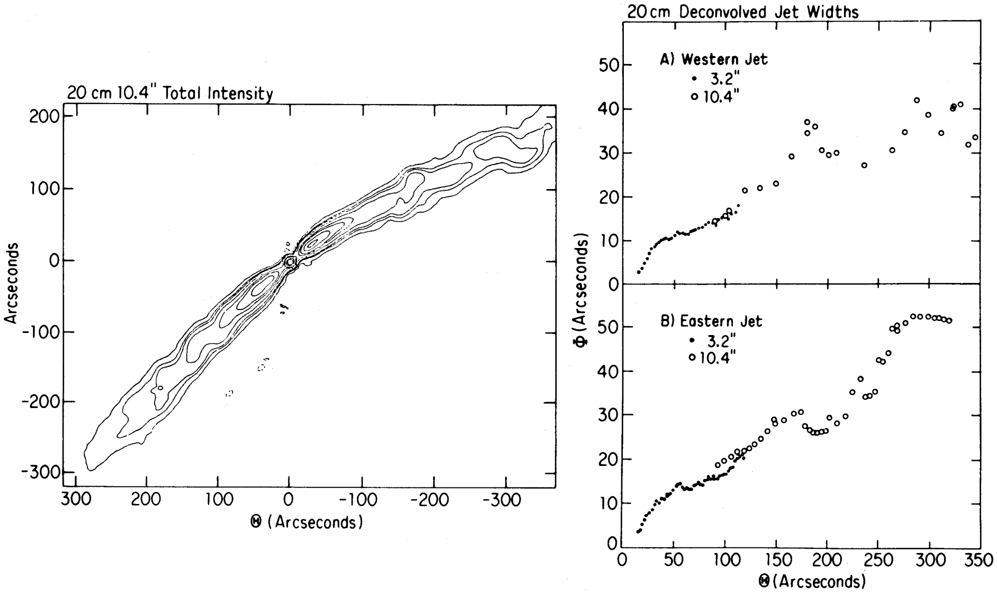


Figure 14–3. An example illustrating the analysis of an image of a radio jet. *(Left)* A contour display of the total intensity. *(Right)* Model-fits of the jet width.

One's comprehension of the properties of a complicated feature is usually enhanced if the dimensionality of the analysis of the feature can be reduced. For example, one-dimensional analyses of filamentary features are useful, and various distributions along lines parallel to and perpendicular to the axes of these features can lend considerable insight. A radio image of a source with a jet is shown in Figure 14–3 (left). A one-dimensional analysis has been made along lines perpendicular to the jet axis at increasing distances from the core (Killeen, Bicknell, & Ekers 1986). The model chosen for the jet emission was a three-dimensional circularly symmetric cylindrical intensity distribution in the plane of the sky. The distribution was Gaussian-shaped, with an unknown peak intensity on the axis, and an unknown width. The best values of the intensity and width were determined from a fit of the model to the image data, after the data had been interpolated along appropriate lines across the jet axis (AIPS task 'SLICE'). The derived width estimates were corrected for the resolution of the image. Other examples are given by Perley, Bridle & Willis (1984). Features with other kinds of symmetry can be analyzed in a similar manner. An example is the determination of the ellipticity of a galaxy image as a function of distance from the nucleus (Killeen, Bicknell, & Ekers 1986, Appendix A).

3.2. The integrated intensity of an extended feature

The integrated intensity, and accompanying error estimate, of an extended feature are important parameters. Because an integrated intensity estimate of an extended feature is obtained from a large area in an image, its value and error are sensitive to a variety of errors in the image. For this reason, a hodgepodge of methods have been suggested for its computation. It is first useful to make a

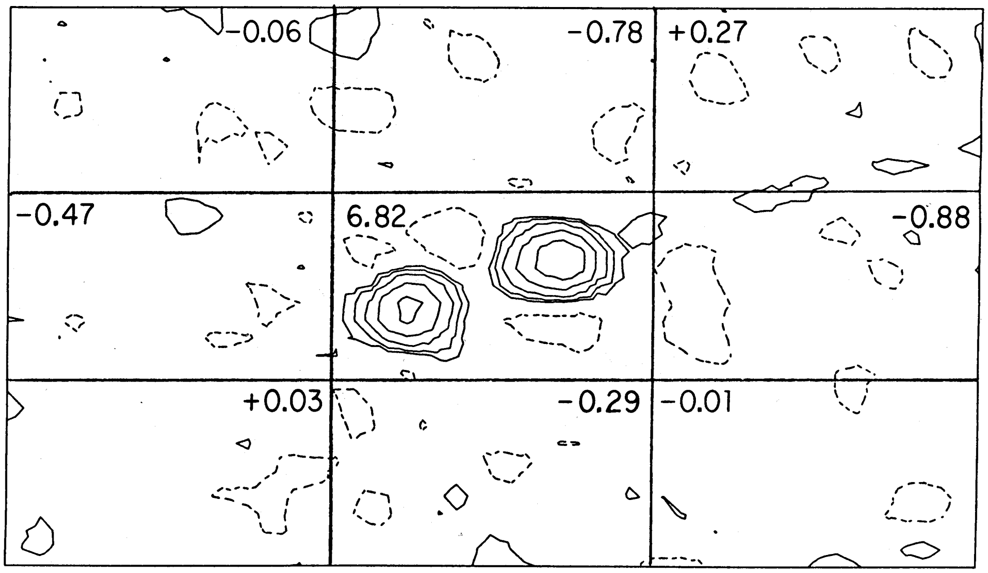


Figure 14–4. The integrated intensity of an extended feature. The peak intensity within the feature is shown at center, as is the peak intensity within each of the eight control regions. This is the same feature that is shown in Figure 14–1.

reasonably low-resolution image in which the feature is not too extended. This will give the best signal-to-noise ratio for the feature, so that optimum component boundaries can be chosen and noise contamination be minimized. The use of several alternative methods for determining the integrated intensity of an extended feature is illustrated below, using the feature in Figure 14–4.

1. Sum up the intensities within the feature as a measure of the integrated intensity (AIPS task 'IMEAN'). A simple summation of the pixel values is usually accurate enough, although a Simpson's-rule integration should be chosen for images with near-critical sampling. Choose several control regions which surround the feature, and use the same integration technique. The integrated intensities in the control regions can be fit to a constant offset, or to a higher-order polynomial 'baseline' around the feature, and an error estimate can be ascertained from this fit. The analysis shown in Figure 14–3, where the control regions were used to solve for a zero-offset and error, gave the result 7.09 ± 0.14. A similar analysis is often used to remove the sky background from an optical image.

2. Fit the feature with a reasonable model of several components (AIPS tasks 'IMFIT' and 'JMFIT'). Do not pay too much attention to the individual parameter values. If the post-fit residuals under the feature are about the same as those in the rest of the field, then the sum of the integrated intensities of each of the components, and its error, should be a reasonable guess for the integrated intensity of the feature and its error. The result for this feature is 8.22 ± 0.45. The model-fit was not particularly satisfactory because of the complicated bias levels around the feature.

3. Sum the 'CLEAN' components within the boundary of the feature (AIPS tasks 'APCLN' and 'IMAGR'). Deep 'CLEAN'ing down to a level of 2 or 3 times the r.m.s. noise may be necessary. No error estimate is given by this method. The value 7.1 was obtained.

Method 1 is preferred, although methods 2 and 3 are satisfactory for features that are somewhat less extended than the one in Figure 14–4. The final estimate depends on the quality of the various methods and their agreement. For this feature a flux density of 7.1 ± 0.2 was used.

3.3. Very extended features

An estimate of the integrated intensity of a very large feature is affected by small biases in the image. Simple sums over the feature can lead to poor estimates, and very low-resolution images often have extremely poor image quality. The integrated intensity over the entire image may be more accurately measured by the intensities of the lowest spatial frequency Fourier components, which can be obtained by a Fourier transform of the image (AIPS task 'FFT') or by direct measurement in aperture synthesis techniques. Extrapolate the lowest frequency Fourier components to zero frequency (AIPS task 'UVPLT') to obtain the integrated intensity F of the image. Such an extrapolation is not always obvious to the eye, but at least some estimate of the value and error can be guessed. At low spatial frequencies the visibility varies as $F - Au^2$, where u is the spatial frequency and A is a constant proportional to the size of the feature. This technique is similar to fitting a model of the feature directly to the visibility data.

4. Image Combination, Analysis, and Errors

This section describes the analyses of a set of images covering one region of sky. The set is most often delineated by frequency, polarization and time, and a host of derived quantities can be calculated from comparison of the intensity distributions in the set. In Section 4.1 I shall discuss how, when and why images are combined, and I shall assume that all of the images have the same resolution and grid frame. More complicated and specialized image combinations, which deal with a set of images at uniformly-spaced frequencies, are described in Lectures 11 and 12.

4.1. Image combination

Given a set of images—all with the same resolution and on the same grid—of the intensity distribution in the sky as a function of polarization ($I_{RR}, I_{RL}, I_{LR}, I_{LL}$), frequency or wavelength (ν or λ), or time (t), many derived properties can be obtained. These calculations are all done on a pixel-by-pixel basis. A list of the more common combinations is:

I	Total Intensity*	$I = (I_{\rm RR} + I_{\rm LL})/2$
m	Magnitude	$m = -2.5 \log I$
I_V	Stokes' V Intensity*	$I_V = (I_{\rm RR} - I_{\rm LL})/2$
I_Q	Stokes' Q Intensity*	$I_Q = (I_{\rm RL} + I_{\rm LR})/2$

I_U	Stokes' U Intensity*	$I_U = i(I_{\rm RL} - I_{\rm LR})/2$
I_P	Linearly Polarized Intensity	$I_P = \sqrt{I_Q^2 + I_U^2}$
ψ	Linear Polarization Position Angle**	$\psi = 0.5 \tan^{-1}(I_U/I_Q)$
I_F	Fractional Linear Polarization**	$I_F = I_P/I$
RM	Rotation Measure**	${\rm RM} = \frac{\psi(\nu_1)-\psi(\nu_2)}{\lambda_1^2-\lambda_2^2}$
D	Depolarization**	$D = I_F(\nu_1)/I_F(\nu_2)$
I_C	Continuum, wide-band*	$I_C = \sum I(\nu_i)$
$I_E(\nu_i)$	Line Emission*, channel i	$I_E(\nu_i) = I(\nu_i) - I_C$
$I_A(\nu_i)$	Line Absorption*, channel i	$I_A(\nu_i) = I_C - I(\nu_i)$
$\tau(\nu_i)$	Opacity**, channel i	$\tau(\nu_i) = \exp(I_A(\nu_i)/I_C)$
α	Spectral Index**	$\alpha = \frac{\log I(\nu_1)/I(\nu_2)}{\log \nu_1/\nu_2}$
$\Delta I(t_1, t_2)$	Variability*	$\Delta I(t_1, t_2) = I(t_1) - I(t_2)\,.$

*Linear combination.
**May have undefined values.

4.2. Linear combinations

The image combinations which are linear (indicated by * in the table) have properties similar to the original intensity distributions: the noise distribution is the appropriate weighted r.m.s. sum of the input images; the noise is nearly spatially invariant over the image; and the point-spread function is identical to that of the input images. All of the image analysis techniques discussed in the previous sections, such as smoothing, interpolating, and parameter fitting, can be applied to the derived images. Since the results of most image reconstruction techniques are linear, images can be combined linearly at several stages of image reconstruction; for example, dirty images can be linearly combined (with suitable weights) before 'CLEAN'ing. The combination of dirty images increases the signal-to-noise ratio and lessens the processing, since the number of images is smaller.

It is even possible to linearly combine the data associated with the images before the imaging algorithms are used. For example, the total-intensity distribution can be obtained from the sum of the visibility data from the two orthogonal polarizations (after calibration!). If the data for the two polarizations are not sampled identically, then some problems may occur in deciding how to combine the data where they are incomplete. There are many subtleties in the imaging process for VLA data, producing small differences between various linear combinations of data which are beyond the scope of this discussion.

4.3. Nonlinear combinations

Many useful astronomical parameters are derived from nonlinear combination of corresponding pixels in the set of images. However, the derived images have properties which must be carefully considered when they are processed and interpreted: the image error distribution is spatially non-uniform and non-Gaussian; an equivalent or effective point-spread function does not exist; the image cannot be smoothed or interpolated; and certain pixel values may be undefined (indicated by ** in the table above). Pixel locations that contain undefined values are indicated by special *blanking* values which are understood by the software reduction system and ignored in subsequent processing and display. For example, in the spectral index calculation (α), any pixel in which either input image contains a non-positive value will be undefined. Images derived from angular quantities (rotation measure, intrinsic magnetic field direction) are ambiguous because of the 180° ambiguity of the linear-polarization angle. Multiple frequency observations, suitably spaced, are necessary to resolve these ambiguities.

Two of the most basic image modifications, smoothing and interpolation, should not be made on images which are nonlinear combinations. The original images (assuming they are linear combinations of the intensity distribution) must be smoothed and interpolated and then recombined. However, if the image is oversampled, then image interpolation of the nonlinear image may be valid in certain restricted regions.

The use of fitting methods to determine properties of an image over an array of many pixels makes little sense for nonlinear images, since the shape of a feature on such an image is poorly defined. For example, how can the spectral index of an extended feature be best obtained? Two methods are available. First, the spectral index image can be formed and then the spectral index, suitably weighted (this may be difficult) over the feature, can be calculated. Second, the integrated intensity and error of the feature can be determined from each of the input images from which the spectral index and its error is derived. The latter method is simpler, but the former gives more information about the spectral index distribution; i.e., is it really constant over the feature? For features that are not too extended, both methods should give about the same results.

The list of image combinations given in the previous table assumes only two input images. Many useful quantities require several input images, and the derived quantities are calculated with algorithms involving all of the images, but always on a pixel-by-pixel basis. There are many specialized programs that deal with a set of images made at a large number of equally-spaced frequencies. See Lectures 17 and 18.

4.4. Image errors

Knowing the properties of errors in an intensity image is essential in determining the reliability of the image, and that of the derived parameters, from image combinations. The errors consist of two components arising from: (1) fundamental limits in the telescope and instrumentation, which produce errors that are stochastic and have reasonably well-defined properties; and (2) systematic effects caused by a variety of instrumental imaging defects, which may or may not be understood, or even suspected.

The gross behavior of stochastic, noise-like, errors depends on the type of receivers and other electronics used in the array. This error is often summarized by one number—*the r.m.s. noise of a pixel.* For correlation-type detectors, as used by most synthesis arrays, the noise is distributed with a normal probability about zero intensity (a small offset is possible) with an r.m.s. dependent on many observing and receiver parameters. Systems with total-power detectors produce noise that obeys a Rayleigh distribution. An example is shown in Figure 14–5. For images on photographic media, the errors scale with the pixel intensity,

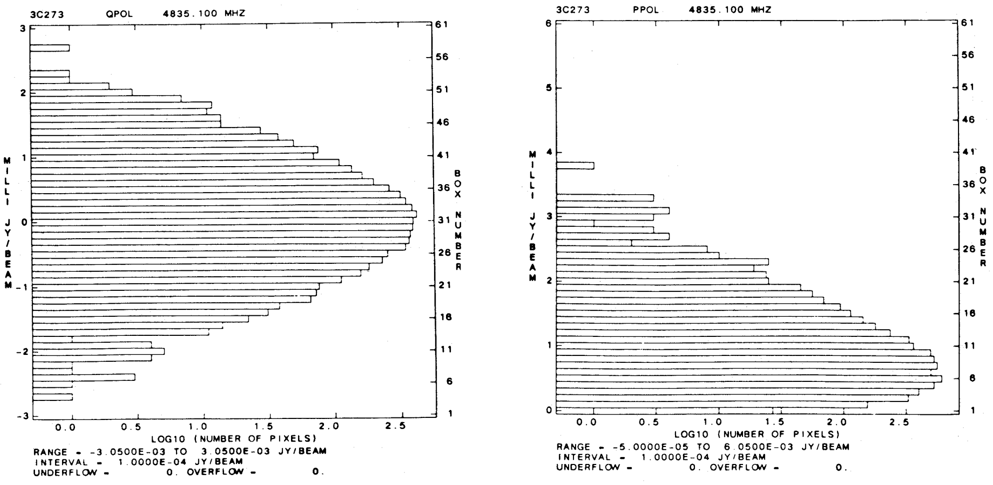


Figure 14–5. The noise distribution within an image. *(Left)* A nearly-Gaussian noise distribution. *(Right)* An approximately-Rayleigh noise distribution.

and the photographic grains produce a Rayleigh-type low-level noise with a characteristic size which is not necessarily the instrumental resolution. Finally, for observations where the number of detected events per pixel is small, the image will have Poisson statistics. These properties are reasonably well-known and can be incorporated in the image analysis.

Systematic errors, caused by a variety of instrumental and imaging problems, cannot be discussed in any detail. They are not as amenable to analysis as the stochastic errors. These errors are serious, in that they are generally different in their spatial correlation length from the random errors. For example, in a VLA image, the stochastic errors between two positions separated by more than one beamwidth are nearly independent. However, systematic errors tend to be larger scale (ripples on an image, zero-level bias) and do not decrease with more averaging. The discussion in Section 3.2 touched on methods to determine the actual errors of parameters in the presence of various types of image problems.

4.5. Errors in image combination

The errors of images made from the fundamental observed intensities, and errors of linear combinations of such images, are summarized by one number—the r.m.s. error per pixel. The correlation length of this error is the beam size, and it is nearly invariant across the image (before correction for the primary beam). This is an approximation, only, and can be misleading if systematic errors are important.

The error analysis of images made by nonlinear combination of input images is not straightforward. The error at any given pixel of the output image can be estimated from the r.m.s. error of the input images. The error distribution may be different from that of the input images, and the output image may have a bias. For example, the linearly-polarized intensity I_P is the r.m.s. sum of the two Stokes intensities I_Q and I_U. If the common noise distribution of the two input images is Gaussian with standard deviation σ, then the noise in I_P will have a Rayleigh distribution, and the image will contain a bias level about equal to the r.m.s. noise (see AIPS task 'COMB', with 'OPCODE'='POLC').

Systematic errors are, of course, harder to deal with because of their unknown properties, and a thorough discussion is beyond the scope of this lecture. Obviously, such errors depend on the instrument. It is most important to try to recognize these errors and to anticipate their effects. For aperture synthesis, Lectures 12 and 15 should be consulted. A different set of problems occurs for optical and X-ray imagery.

In addition, systematic errors may be fully, only partially, or un-correlated between the input images, and the induced error in the output image will depend on the characteristics and on the origin of the error. For example, the circularly polarized intensity distribution will not contain errors caused by tropospheric pathlength errors, since both orthogonal polarizations are equally affected and cancel when the two orthogonal polarization images are differenced. Thus, the quality of this image is a good indicator of data errors not due to propagation effects.

Two methods are used to indicate the error at each pixel within an image constructed by combination of other images. The most natural method is to calculate an image containing the estimated error at each pixel. Those pixels which are undefined can be set equal to the *blanking* value. Such error estimates can be accurately obtained for the stochastic error contribution. The systematic errors associated with imperfections in the image—for example, low-level intensity waves associated with extended features—are difficult to quantify, and thus the trustworthiness of derived parameters can be hard to assess.

A more compact method of error indication, which avoids the generation of an additional error image, is to assign any pixel in the derived image the special blanking value if its estimated error exceeds a certain value. Further analysis and display algorithms will ignore the blanked pixels. Often, the best blanking level is not known *a priori*, so that the image combination must be repeated several times to give the desired output. The valid pixels all have equal weight, so that further averaging or reliability judgements are difficult to make.

4.6. Image alignment

Images must be aligned before they can be combined and compared. A set of images which have been obtained simultaneously from observations is usually precisely registered, although some processing may shift one image with respect to another. For observations made at different epochs, even using the same instrument in an identical configuration with identical calibration strategy, offsets between the images occur because of inaccuracy in the determination of the absolute positioning. In single telescopes, systematic errors in the pointing between observations cause relatively large registration errors. For synthesis arrays, registration errors are produced by errors in the positions of the antennas, offsets in the time-keeping, and inaccuracies in the model for removing atmospheric and ionospheric refraction. These registration problems can be minimized somewhat, via proper calibration (see Lecture 5) and by careful monitoring of the instrument while the observations are recorded. The intensity scales between several images may also differ.

The final adjustment of the registration of a set of images is often accomplished in the *ad hoc* manner of aligning bright, unresolved features for which there is external evidence that these features are coincident. For radio–optical comparison of images, registration errors can be as large as $1''$, and better alignment is obtained by assuming that compact radio and optical components are coincident. For VLBI observations, where resolutions are much higher than the absolute positioning, proper alignment of images obtained at different frequencies or at different epochs is difficult to achieve.

The coordinate system of many images which cover a large field-of-view is not precisely linear. The nonlinearities are caused by a variety of effects. Some examples are:

1. Changing nonlinearity for non- East–West aperture synthesis arrays (Lecture 15).
2. Different projection of the sky by various instruments (Greisen 1983).
3. Misalignment of arrays of detectors.
4. Mosaicing of adjacent images (Lecture 20).

The forms of these distortions are generally known, and they can be corrected by using Equation 14–1 in order to interpolate all of the images onto a common grid.

4.7. Other image combinations

Most image combinations like those discussed above make little sense if the input images are not of identical resolution. Such combination will produce strange effects near the edges of discrete features, and the intensity scales will not be directly comparable. The resolutions of the images can be equalized using the techniques of Section 1.

Images with different resolution, but which are otherwise identical, can be added together. However, the corresponding sum of the point-spread functions must also be calculated in order to interpret these images. For example, if a radio source is observed for many consecutive days, then it may difficult to store all of the visibility data in the data reduction system. The sum of the day-by-day images and beams produces an image which is nearly as accurate and sensitive as that obtained from imaging the entire set if all days' observations are nearly identical. Dirty images and beams from observations at different resolutions (configurations) can be summed and then 'CLEAN'ed. However, there is much more control over the relative weighting of the visibility data if the data are combined before imaging.

The usual theory of aperture synthesis assumes that the intensity distribution does not change during the time-period of the observation. If the distribution does vary (or apparently vary because of instrumental problems like the rotation of the primary beam on the sky), then image reconstruction of the entire set of data will not produce the best-quality images. For example, if a point component in the image has significantly varied over the observing period, then the best 'CLEAN'ed and self-calibrated images will still contain artifacts from the variable source. The simplest solution is to split the observations into segments during which the variability is not significant, obtain the best-quality images, and then sum them. Other, more sophisticated methods are discussed in the next section.

5. Selected Image Analysis Topics

This final section includes a *potpourri* of topics important in the analysis of images, especially those obtained from VLA observations. This list is not exhaustive—much of the discussion is experimental and is meant to foster further debate.

5.1. Absolute positions

The accuracy in determining the absolute position of a radio source is largely controlled by the weather, the signal-to-noise ratio of the observations, and the accuracy of the assumed positions of the sources that were used to calibrate the data. For weak sources, the expected positional accuracy is given in Section 2.3. If this error is larger than nominal calibration errors ($\sim 0''.1$ for **A**-configuration), normal observing and imaging techniques should then be used. For strong sources, which will be dominated by systematic effects, suggested observing strategies include the following:

1. Observations for one hour of about 10 calibrators spread over the sky, in order to determine residual baseline and timing errors.
2. Observations of a calibrator with an accurately known position, within about 10° of the source.
3. Observations of several other calibrators which bracket the source.
4. Avoid observing at low elevation angles.
5. And pray for good weather.

After proper calibration of the data (baselines and temporal phase variations), determine the position derived from each observation of the source and the calibrators using an image fitting routine ('IMFIT' or 'JMFIT'). From this ensemble of positions for the calibrators and source, it should be possible to determine an accurate position and a realistic error estimate for the source. A rough estimate of any systematic effects can be obtained from the positions of the calibrators which bracket the source.

For observations at low frequency when the dispersive ionospheric refraction is large, simultaneous dual-frequency observations can remove the effects of this refraction and improve the accuracy of the position determination. Suppose a field is observed simultaneously at 1.4 GHz and 1.6 GHz. If the dispersive refraction at any instant is R at 1.4 GHz, the refraction at 1.6 GHz is then $(1.4/1.6)^2 = 0.77\ R$. The difference in position between the two frequencies is thus a measure of R. The radio image of the source must be made over sufficiently short time-scales during which the refraction is relatively constant. If the radio field is dominated by one strong source, the observed visibility phases between the two frequencies can be combined to remove the ionospheric refraction (see Lecture 5, Sec. 10.3).

5.2. Relative positions and motions

The accurate motion of a radio component over a period of time is best measured by comparison of its position with that of other radio sources in the same radio image. The accuracy here is limited by signal-to-noise considerations instead of weather, and by calibrator positional accuracy for the absolute position. Depending on the signal-to-noise of the observations, several schemes are available. In all cases, observations of the radio field should be made with the same array configuration and a similar observing schedule. This is crucial for the high dynamic range images and convenient for the lower signal-to-noise images.

For the lower dynamic range images ($500:1$ or less), image, 'CLEAN' and self-calibrate the field in the usual way. Make sure the same 'CLEAN' beam is used for all of the images. Measure the position of all radio components (AIPS task 'IMFIT' or 'JMFIT'). Analyze the differences of the positions of all sources in the image to determine any relative motion between the epochs. If some of the radio features are extended, then an angular cross-correlation between the two images can be used to determine the offset of the feature between two epochs.

For high dynamic range imaging, subtle effects in the image reconstruction algorithms can introduce slight differences in the images, which can be amplified and result in bogus differences in subsequent processing. It is better to difference the visibility data in some manner, between the two epochs, before any processing. Many schemes are available, and it unclear which are the best.

When comparing the relative positions of sources over a region larger than about $5'$, small relative distortions and rotations occur at the $0\rlap{.}''1$ level. These are produced by differential aberration, precession and nutation across the image—which depend on time. Exact details of their removal depend on precession implementation in the on-line system (and the use of 'UVFIX' in AIPS). The algorithms are given in the standard ephemeris.

5.3. Intensity variations

Intensity variations of radio features between epochs can be determined in the usual manner. Intensity variations during an observation, however, are more painful to sort out. As described in the previous section, such variations produce artifacts which cannot be 'CLEAN'ed and self-calibrated away. The defective images contain artifacts which are generally symmetric about the varying object. A useful method for determining the intensity variations, and removing them, is as follows:

1. Split the observations into N segments. Something like $N = 4$ might be a reasonable first try.

2. Image, 'CLEAN' and self-calibrate (phase only) each segment.

3. Measure the flux density of the variable source.

4. Subtract the visibility function of the variable part of this source from each of the segmented databases.

5. Image, 'CLEAN' and self-calibrate the entire data set. The steady, base-level flux-density of the variable component should remain.

If this segmented approach produces a better-quality image, then further segmenting may also help.

References

Greisen, E. W. 1983, AIPS Memorandum No. 27, NRAO.

Killeen, N. E. B., Bicknell, G. V., & Ekers, R. D. 1986, *ApJ*, 302, 306–336.

Perley, R. A., Bridle, A. H., & Willis, A. G. 1984, *ApJS*, 54, 291–334.

Weast, R. C. & Selby, S. M. 1975, *Handbook of Tables for Mathematics*, Revised Fourth Edition, CRC Press (Boca Raton, FL), p. 865.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

15. Error Recognition

R .D. Ekers
Australia Telescope National Facility, Epping, NSW 2121, Australia

Abstract.
Ways to recognize several common image defects are presented. Some useful diagnostic tools are also discussed.

1. Introduction

In this lecture I have two main aims: to use the discussion of image defects to give a better feel for how you can understand a synthesis telescope such as the VLA with the aid of a few basic concepts and analogies, and to provide some practical information for use when observing with synthesis radio telescopes.

Most of the image defects are caused by errors which occur in the measurement plane (i.e., the (u, v) plane). In synthesis imaging, the image in the (l,m) plane is the Fourier transform of the visibility data in the (u, v) plane. But it is the effects of the errors in the image plane that finally matter, so we must make heavy use of the relationships between the two Fourier domains. The collection of Fourier transform pairs in Bracewell (1978) provides an excellent source of inspiration when considering the relations between the two domains. Since the sky brightness takes on only real values, the data in the (u, v) plane must be Hermitian, so instead of measuring visibilities over the whole (u, v) plane we fill in half of it with the complex conjugates of the values measured (with the baseline orientation reversed) in the other half of the plane. Because of this, we need handle only Fourier transform relationships between Hermitian functions and real functions.

2. Diagnosing Errors

2.1. Image plane or (u, v) plane?

We have two contradictory requirements: Since the errors usually occur in the (u, v) plane they are often more readily recognizable and easier to diagnose in the (u, v) plane, but it is their effects in the image plane that finally matter—so, unless the effects are important in the image, there is no point in diagnosing them! The Fourier transform of a serious error may not be so serious. For example, we can totally destroy a small part of a hologram, with little effect on the image generated from it. Of course this is just the reason why we can succeed in making a reasonable quality radio image even when we haven't measured over all of the (u, v) plane. The holes in the (u, v) sampling with completely incorrect values (zeros for the principal solution) don't completely destroy the image.

One of the most important of the Fourier transform properties is that a sharp peak in one domain transforms to a broad feature in the other (Fig. 15–1a or 15–1d). Consider the effect of a single bad value in the (u, v) plane. We put the complex conjugate of this value in the opposite half of the (u, v) plane and make the image. The situation then corresponds to Figure 15–1b, and

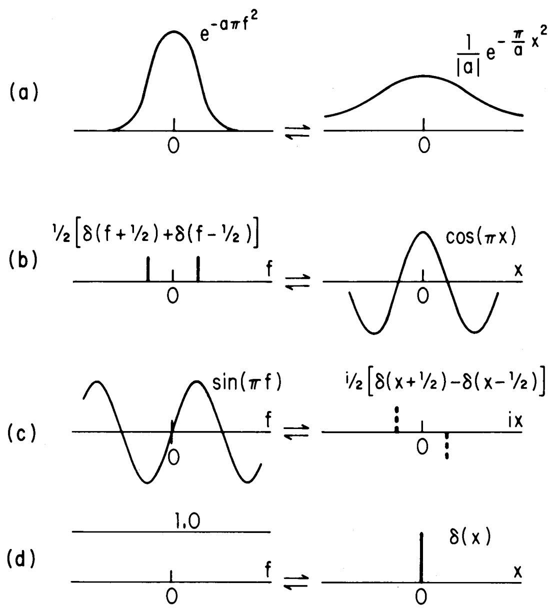

(a)
$e^{-a\pi f^2}$
0
$\frac{1}{|a|} e^{-\frac{\pi}{a} x^2}$
0
(b)
$\frac{1}{2}[\delta(f+\frac{1}{2})+\delta(f-\frac{1}{2})]$
f
0
$\cos(\pi x)$
x
0
(c)
$\sin(\pi f)$
f
0
$i\frac{1}{2}[\delta(x+\frac{1}{2})-\delta(x-\frac{1}{2})]$
ix
0
(d)
1.0
f
0
$\delta(x)$
x
0

Figure 15–1. Fourier transform pairs. The broken lines indicate the imaginary domain. For many more examples see Bracewell (1978).

the transform of the pair of error delta functions produces a sinusoidal ripple through the image. The effect of this error is then spread over the entire image, so the relative amplitude of the erroneous sine wave in the image will be very much smaller than the relative amplitude of the erroneous point in the (u, v) plane. For example, consider an observation of a point source of flux density S. At the position of the peak of the source in the image, the Fourier transform will correspond to the sum of all the observed visibility samples. The number of samples in a full observation with the VLA is $N = 351$ (interferometers) $\times$ 2880 samples (8 hours with 10 – second sampling) $\approx 10^6$. Any point in the image is a linear combination of these N samples, giving an amplitude S; hence each sample has weight $1/N$ (assuming natural weighting). A single erroneous visibility with amplitude ϵ will cause an error sinusoid with peak amplitude ϵ/N. If we want an image with peak error $< 0.1\%$ then $\epsilon/N < 10^{-3}S$, so for $N = 10^6$ we need only remove errors with amplitude $\epsilon > 10^3 S$! This illustrates that, whereas an error of this kind would be easily detected in the (u, v) plane, there is very little point in doing so unless the error is thousands of times larger than the correct value. If the observation is much shorter, a single erroneous point will have more effect. For the above example in the case of a 30-sec "snapshot" observation $N = 10^3$, so a single bad value of the amplitude S will be important at the 0.1% level.

Compare this to the situation with an error which is very spread out in the (u, v) plane. For example, consider an error caused by one correlator having a constant offset for the entire observation. Near the center of the final image all the affected (u, v) points will add with the same phase; this is 2880 (samples) times worse than the case we first considered of a single bad point. Summarizing, the errors which are easy to detect in the (u, v) plane must have very large amplitude to be important, but some of the subtle effects in the (u, v) plane can cause bad errors in the image plane—so that is often the best place to look for them.

2.2. Short and long time-scale errors

Short time-scale errors in the (u, v) plane produce large angular scale features in the image, whereas long time-scale errors in the (u, v) plane will give small angular scale effects in the image plane. In the normal two-dimensional situation the error often has a large scale in one direction and a short scale in the other direction. For example, if a single interferometer has an error which is fairly constant in time, this will be a slowly varying error along the direction of the (u, v) track, but a very sharp error in the direction normal to the (u, v) track. The corresponding error in the image plane will have a small angular scale in one direction and large angular scale in the perpendicular direction. Typically this will result in a corrugation in the image plane. The rate at which the error corrugation falls off with distance depends on the duration of the error. If only a single point is wrong the corrugation will have constant amplitude over the entire image, but if a whole scan is wrong the corrugation will die away on a scale which is inversely proportional to the scan length.

Another important clue results from the geometry of the Earth-rotation synthesis. If an error has a long duration and the source declination is not near 0° the error will cause a ring of anomalous values (or a segment of such a ring) in the (u, v) plane; this transforms into a feature in the image whose radial profile resembles a Bessel function, producing an obvious concentric ring structure (e.g., Fig. 15–2). However if the error has occurred for a very short duration, the anomaly will not be a ringlike feature in the (u, v) plane, and its transform will contain only linear features (e.g., Fig. 15–3a or b).

2.3. General forms of errors

The errors $\epsilon(u, v)$ can be divided into different types, depending on whether they correspond to modifications of the visibility data $V(u, v)$ that are additive, multiplicative, convolutions with other functions, or more complex corruptions.

Additive errors are those whose Fourier transform $\mathfrak{F}\epsilon$ is added to the image and is independent of the position and amplitude of any other structure in the image, i.e., for which

$$V + \epsilon \rightleftharpoons I + \mathfrak{F}\epsilon\,; \qquad (15\text{–}1)$$

where the $\rightleftharpoons$ symbol here denotes a Fourier transform pair relationship between quantities in the measurement $((u, v))$ plane (left-hand side) and in the image $((l,m))$ plane (right-hand side). Examples of additive errors are interference, cross-talk, correlator offsets, and receiver noise.

Multiplicative errors are those for which

$$V\epsilon \rightleftharpoons I * \mathfrak{F}\epsilon\,; \qquad (15\text{–}2)$$

i.e., the Fourier transform of the error is convolved with the source distribution in the image. Examples are atmospheric and ionospheric phase errors, calibration errors in amplitude or phase, and multiplicative baseline-based errors (closure errors).

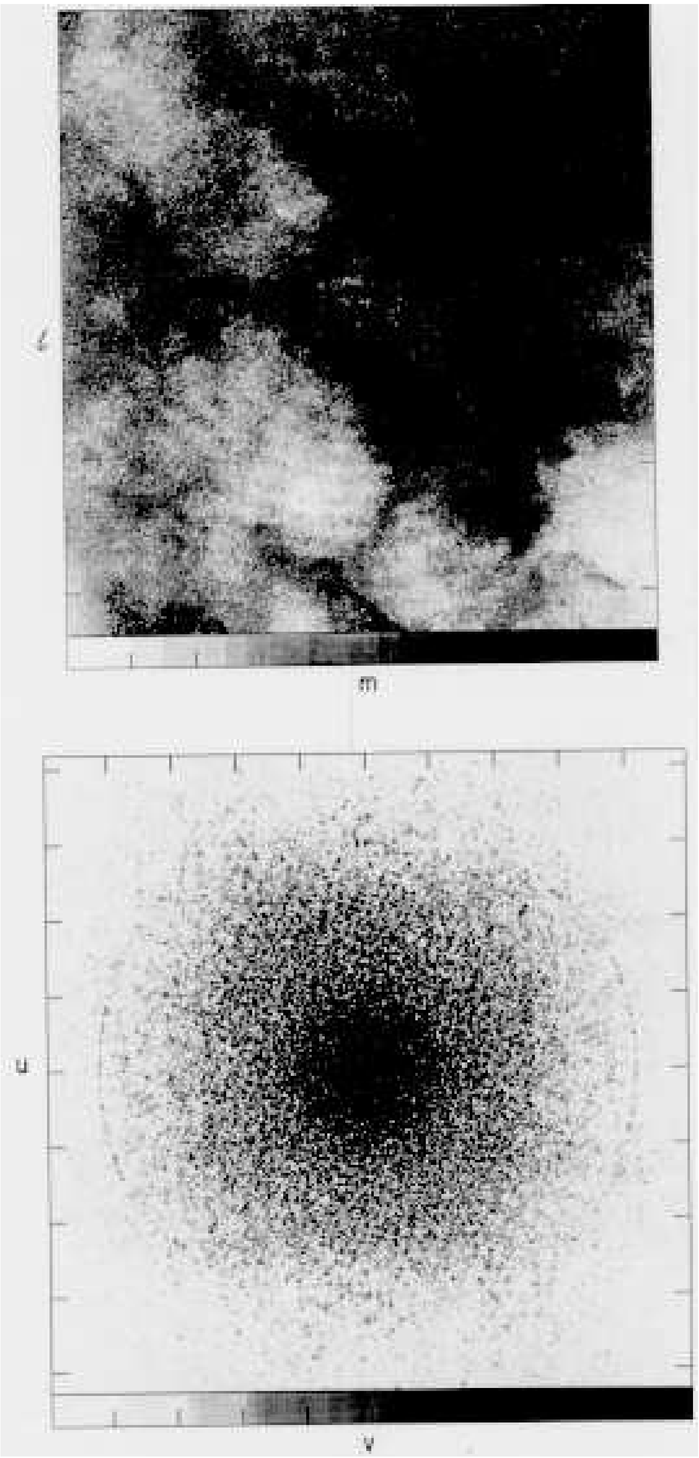


Figure 15–2. *(Top)* The inner portion of an image of Cassiopeia A, centered on the phase-tracking center. A baseline-based error which persisted throughout the observations caused the concentric rings in this image. *(Bottom)* The (inverse) Fourier transform of the above image. The two curved linear features near the left and right edges of the display correspond to the locus of the error-corrupted interferometer baseline.

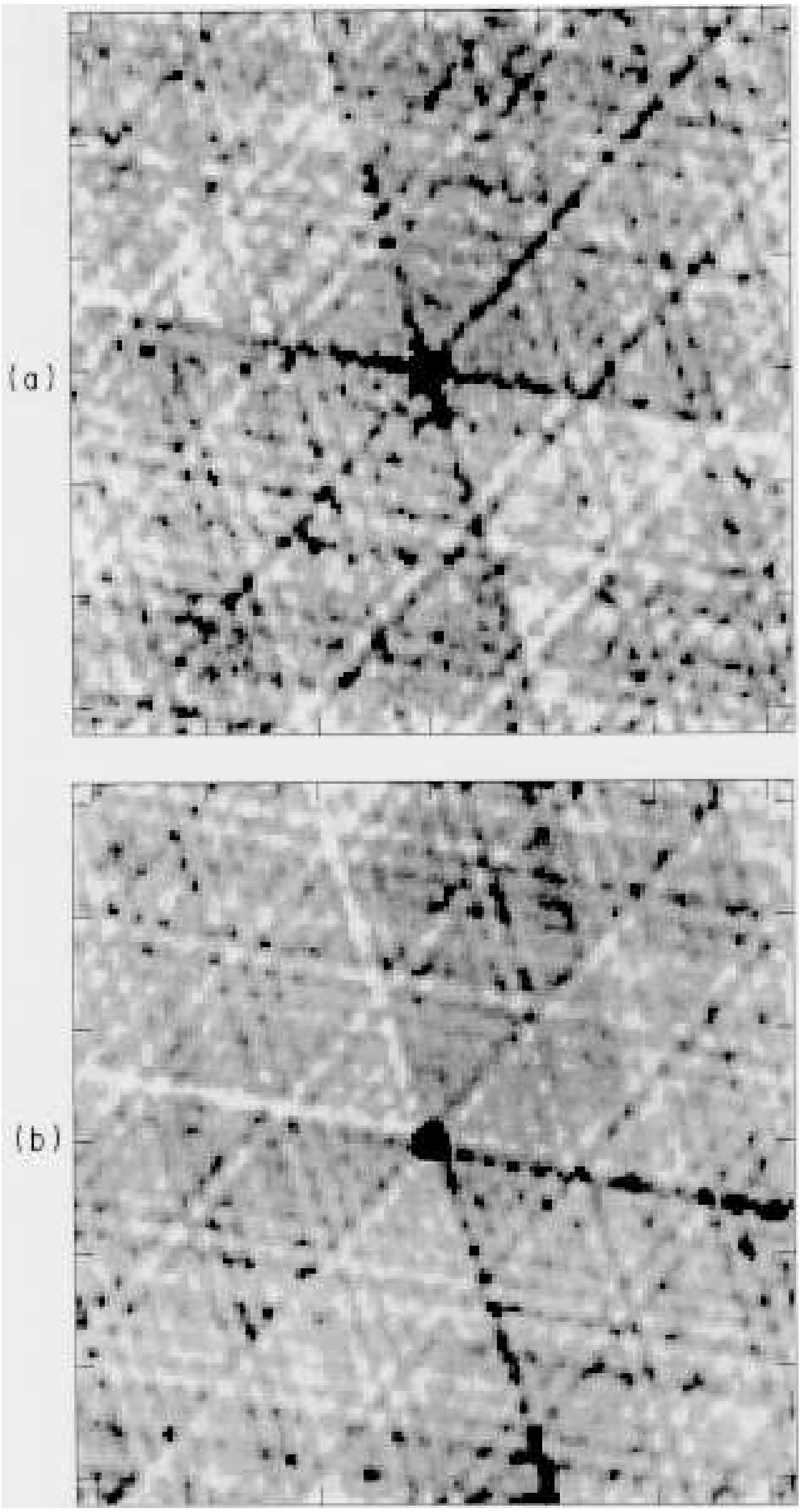


Figure 15–3. Images from a "snapshot" observation of a point source **(a)** with a 10% amplitude error on one antenna and **(b)** with a 10% phase error on one antenna.

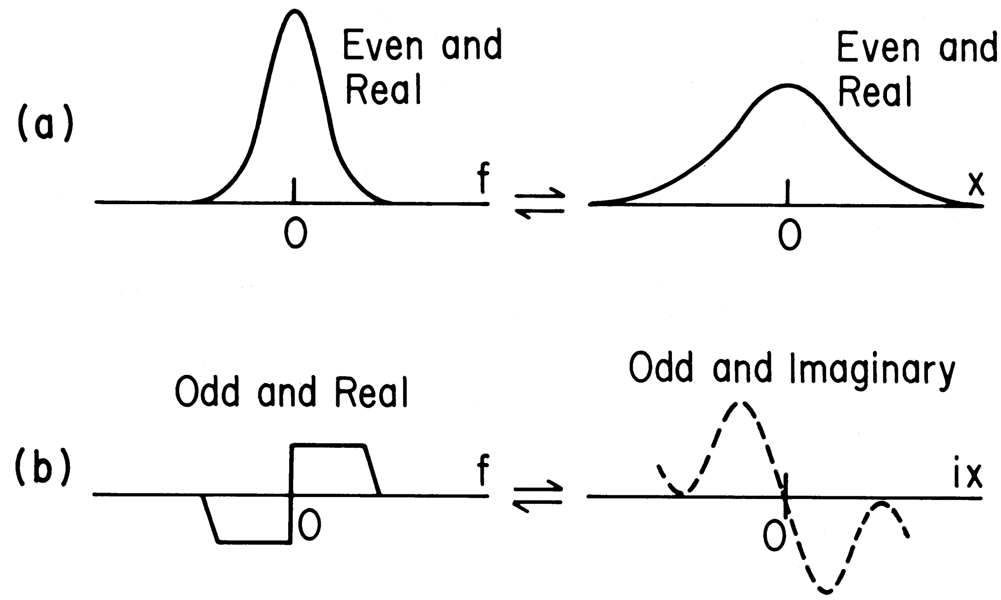


Figure 15–4. Fourier transforms of symmetric and asymmetric functions. The broken line indicates the imaginary domain.

For errors corresponding to a *convolution* of the observed visibility function we have

$$V * \epsilon \rightleftharpoons I \mathfrak{F}\epsilon\,, \tag{15–3}$$

so in this case the image is multiplied by the Fourier transform of the error function. Examples are the effect of the primary beam of the array elements and the convolution needed to resample for the fast Fourier transform (Lecture 7).

Finally, there are errors which are like a convolution in the image plane but which increase in severity with distance from the phase center, delay center or pointing center for the observations. For example, bandwidth smearing (Lecture 18) is a multiplicative error in the (u, v) plane that depends on the source position, so in the image plane it becomes a spatially dependent smearing, rather than a simple position-independent convolution. Other examples are time-average smearing, baseline errors, pointing errors, and shadowing errors.

2.4. Real and imaginary parts of errors

If the error term $\epsilon(u, v)$ is real-valued, then, since the Fourier transform of an even, real function must be symmetric (Fig. 15–4a), this will produce a symmetric error pattern $\mathfrak{F}\epsilon$ in the image. If the error term has an imaginary component, then the Fourier transform of this imaginary odd quantity will give an asymmetric component to the error (Fig. 15–4b) in the image. Hence, by looking at the symmetry, or asymmetry, of the error pattern in the image plane it is possible to tell whether the cause is a real or an imaginary error in the (u, v) plane. This difference is illustrated for a short VLA "snapshot" observation in Figures 15–3a and 15–3b.

3. Examples

3.1. Additive errors

These are errors of the form given by Expression 15–1. They result in an error pattern which is added to the image and is unrelated to the amplitude or position of any features in the image.

3.1.1. The Sun Sources of radiation which are far away from the position being observed are suppressed by the primary beam of the array elements, by the sidelobes of the synthesized beam, and by the bandwidth smearing described in Lectures 2 and 18. However, the solar emission can be 10^{11} times the level being studied in the image, so it may not be adequately suppressed, even if the Sun is tens of degrees away. Since the Sun has a relatively large angular size, and since the bandwidth smearing selectively suppresses responses from the longer spacings, the errors in the image which are caused by the Sun will be very broad. The effects of solar interference will therefore be very much worse on narrow-bandwidth observations, or on observations using compact arrays. One way to check whether the error has been caused by the Sun is to look at an affected baseline in the (u, v) plane. Since the Sun is likely to be a long way from the position of the observation it will cause rapid variations which can by seen by plotting the visibility as a function of time. The approximate angular distance to the source of the interfering signal can be calculated from the period of oscillation in the visibility function. This is also an example of an error which will look very severe in the (u, v) plane but, because the variations are very rapid, their effect on the image may not be important.

3.1.2. Interference Interfering signals have two properties which are important in determining the nature of their effects on an image. They may fluctuate in intensity (or have a very short duration), in which case they will transform to features which cover a large angular scale in the image. If the interference is occurring on large baselines, the features will have a small fringe spacing even though they are spread over a large scale. Secondly, they will be coming from the wrong direction and will not be moving at the sidereal rate. This means that they will only produce a strong response on baselines for which the expected fringe rate is near zero. The example in Figure 15–5, taken from Thompson (1982a), shows the result of a constant source of interference on an 8-hour observation. The interference has caused horizontal stripes through the image because the only baselines for which the expected fringe rate is zero are those which project to a North–South orientation.

Another way to look at this is to note that a source of interference at a *fixed* position is like a source at the North Pole. Hence the pattern of horizontal stripes is just a small section of a set of rings concentric with the North Pole.

3.1.3. Cross-talk This is the same kind of effect as external interference, except that the interfering signal is generated in one antenna and transmitted to another. Since it usually occurs between close antennas, it is a more serious problem in compact arrays (such as the **D** configuration of the VLA), and it results in an error in the image with a very large angular scale.

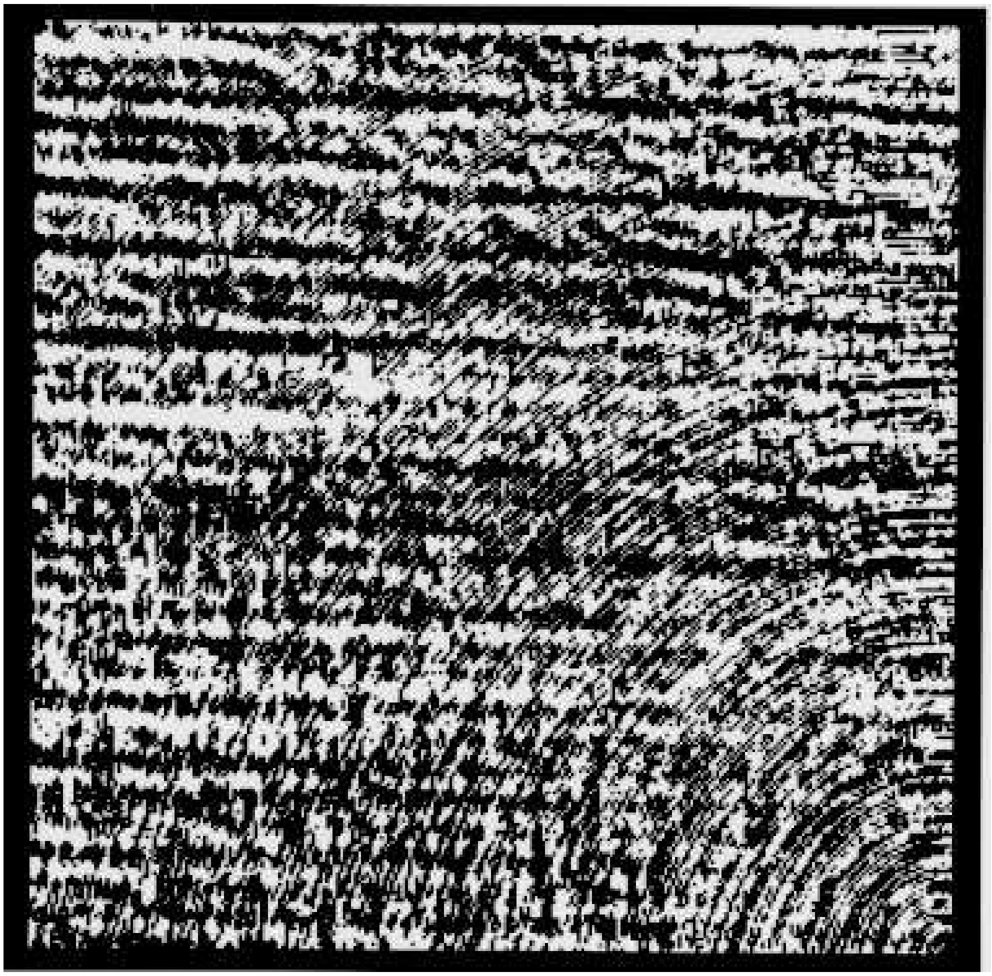


Figure 15–5. Top left quadrant of an image of a point source (bottom right corner) with continuous narrow-band interference. [From Thompson (© 1982 IEEE).]

3.1.4. Baseline-dependent errors Baseline-dependent errors (such as offsets in the correlator) affect individual interferometer baselines. They may take the form of a single bad data point, as was discussed in Section 2.1, or of small constant offsets for the entire observation. A constant offset in the data for one baseline is identical to the response produced by a point source at the phase reference position used by the on-line computer (Lecture 2). Hence, if all the baselines had the same constant offset, the result would be indistinguishable from a point source at the phase reference position. In practice, the offsets will vary from baseline to baseline, so that the result will be an error with absolute maximum amplitude near the reference position and with a sidelobe pattern determined by the distribution of offsets. Furthermore, the time-varying calibration of the atmospheric phase errors will redistribute the phase of the error, reducing the effect on the image. For this reason, baseline-based offsets are less important in higher-resolution observations. Since there are separate correlators for each polarization, the error is likely to be highly polarized.

Although such errors are kept to a very low value in the VLA, they are not completely unknown there. Measurements of very weak sources or point source detection experiments will be more reliable if the phase reference position is displaced a few beamwidths from the position of the object of interest.

3.1.5. Noise This form of error has been extensively discussed in Lecture 9. One additional point may be of interest. Since the receiver noise only occurs at places in the (u, v) plane where the visibility has been measured, it will appear to have the same sidelobe structure in the image plane as would a real source.[1] Consequently, the presence of sidelobes does not provide a method of distinguishing between a real source and a noise fluctuation in the image. In images made from data with well-filled arrays, the peak sidelobe level will be low enough that this effect will be noticed only for noise fluctuations that are well above the r.m.s. noise (e.g., $\geq 5 \times$ r.m.s. for VLA data). Note however that such fluctuations are not unlikely in large ($>$ 1000 pixel) images! The effect is most noticeable in images from telescopes, such as Westerbork or Fleurs, that produce strong grating responses.

3.2. Multiplicative errors

These are errors of the form of Expression 15–2. Since they result in a convolution in the image plane they appear to be "attached" to the sources in the image.

3.2.1. (u, v) coverage effects A serious "error" in our data is caused by all the missing information in the (u, v) plane. Where the data are missing, the source visibility $V(u, v)$ has effectively been multiplied by zero. This is not usually called an error and, as discussed in Lecture 8, we normally attempt to correct its effects by using some form of deconvolution algorithm, e.g., 'CLEAN'. How well we do depends on the size of the unsampled regions and their location relative to significant structure in the visibility function, especially near $u = v = 0$. The problem is that when you look at your raw image it is difficult to distinguish the effects caused by the missing information from effects caused by errors in the measured data. Our main clue about the nature of effects caused by the missing information is in the point spread function (dirty beam). The sidelobes of this dirty beam are the (negative) Fourier transform of the missing information. If an image has features around the sources which look just like the sidelobe pattern of the dirty beam then this is most likely to be an effect of the missing (u, v) spacings. If we see effects which have a very different shape then they *may* be caused by errors in the data. But beware of the following complication: When making this assessment we use the dirty beam to give us a way to gauge the effect of the missing information on a *point source.* The sidelobe pattern for an *extended source* is not the same. A very extended source is affected only by the information that is missing at *short* (u, v) spacings. Although this information is included in the point source response function, it may be present with such low

[1] This effect will be most noticeable on "dirty" images. Deconvolution will redistribute the errors, incidentally making false sources produced by noise "spikes" seem more convincing! — *Eds.*

amplitude that it is completely masked by higher amplitude sidelobes coming from the missing information at large spacings. Thus, even if your image shows large amplitude, broad sidelobes that do not seem to be present in the dirty beam, these sidelobes may still be caused by poor (u,v) coverage. To find out whether they are caused by (u,v) coverage you could either make an image and its beam with a taper chosen to emphasize the scale of the broad structure, or see whether the putative sidelobes are removed by a deconvolution algorithm.

3.2.2. Gain calibration errors The problems of calibrating the amplitude and phase (complex gain) of each antenna were discussed in Lecture 5. Any errors introduced as a result of this calibration multiply the visibility function, so their effect on the image is to convolve each source with the Fourier transform of the calibration error. Amplitude calibration errors, i.e. $\epsilon(u,v)$ real, give rise to symmetric error patterns associated with each source in the image. Phase calibration errors, i.e. $\epsilon(u,v)$ imaginary, give rise to asymmetric patterns, as discussed in Section 2.4 above. Figure 15–3a shows the effect of an amplitude calibration error, and Figure 15–3b the effect of a phase calibration error.

For the VLA, the amplitude and phase calibration is antenna-based, so any error will affect all interferometers involving that antenna. In a long observation the Fourier transform of an error confined to this set of interferometer tracks will have a ring-like structure (as in Fig. 15–2). This ring-like structure degenerates to a linear structure near 0° declination. In a short VLA observation the distribution of all interferometers associated with one antenna is a “Y”—so an antenna-based calibration error produces an artifact in the image that looks like a six-pointed star associated with every source (as in Figs. 15–3a and 15–3b).

3.2.3. Atmospheric (and ionospheric) errors Differences in the refractive index of the atmosphere along the line-of-sight from the different antennas to the radio source cause phase differences which do not correspond to source structure. The magnitude of the atmospheric phase error increases linearly with increasing spacing up to a few km, and then the fluctuations become uncorrelated. In the linear regime (i.e., the **D** and **C** configurations) if we have a phase difference of Δ radians per wavelength of baseline, the visibility is modified to

$$V(u)e^{-2\pi iu\Delta}\,, \tag{15–4}$$

and then by the shift theorem for Fourier transforms we have

$$V(u)e^{-2\pi iu\Delta} \rightleftharpoons I(l-\Delta)\,. \tag{15–5}$$

Hence, the effect of an atmospheric phase error is to shift the position of the source. For longer baselines, a random phase error will be introduced; this will cause the image to be convolved with an asymmetric error function (see the example in Fig. 15–6). If the fluctuations occur on a short time scale compared with the length of the observation, then the resulting image will be smeared out and will have reduced amplitude. This is equivalent to “bad seeing” in optical astronomy, with one important difference—in the optical case the aperture is always filled, so all of the spatial Fourier components are measured at all instants in time, and the smeared-out image is the superposition of many perfect instantaneous images (*speckles*) which dance around in time. In the

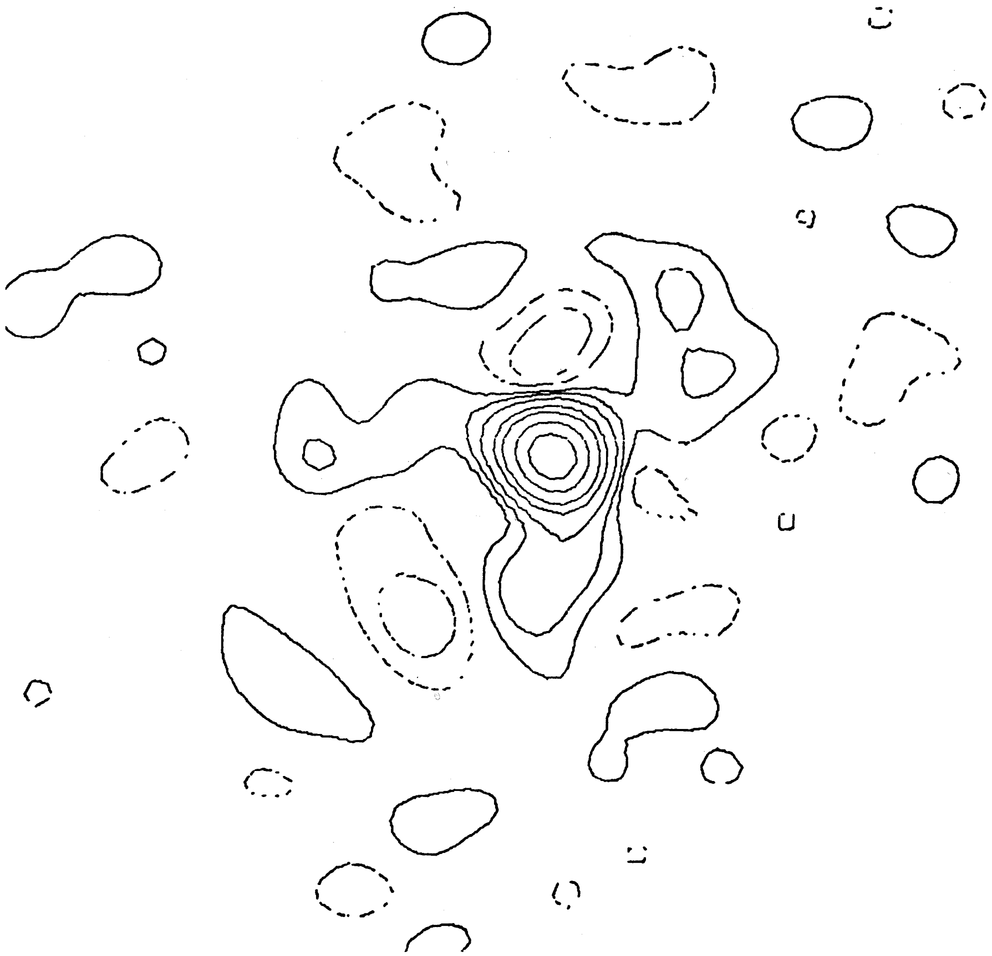


Figure 15–6. Asymmetric pattern which could be caused by atmospheric phase errors.

synthesis telescope, each instantaneous image has a different sidelobe pattern. Consequently, the feature in the final image is not only smeared and reduced in amplitude, but it also has a higher than average sidelobe pattern which can be spread over a large area. Since the atmospheric errors are antenna-based they can be removed by the self-calibration technique described in Lecture 10.

If the atmospheric effects have a long time scale compared with the length of the observation, the result is a good image but one which may be displaced from the correct position. This can occur in compact arrays when the atmosphere above the telescope contains a wedge (perhaps a slowly moving weather front) which remains constant for the observation but is not completely removed by correcting for the phase gradient observed for the calibrator. In this case the resulting image will appear to have high quality, but the sources will be displaced from their correct positions. When this situation occurs, the combination of short "snapshot" observations made at different times may result in a worse image than that from any of the individual "snapshots".

3.3. Errors increasing with distance from the phase reference center

In general these errors cannot be expressed as a simple operation in the image plane, but, if an error has a linear dependence on the radial distance from the image center, then it can be corrected to the form of Expression 15–2 by converting to exponential radial coordinates, as discussed in Lecture 8 of the 1982 *Synthesis Mapping Workshop* (Section 3, Eqs. 8–11 and 8–13).

3.3.1. Bandwidth and time-average smearing These effects are discussed extensively in Lecture 18. Their characteristics are easy to recognize in an image, since the nonzero bandwidth produces a radial smearing, and the time constant causes an approximately tangential smearing. The bandwidth effect is like adding together images with different angular scaling corresponding to the range of frequencies in the band. At the North Pole the averaging-time effect is exactly like a rotational smearing corresponding to the range of times in one sample. Away from the North Pole, different baselines are smeared by different amounts, giving a more complicated result. Both these effects increase monotonically with the distance from the center of the image.

3.3.2. Shadowing of the antennas At low elevations and in compact arrays, it is possible for one antenna to be blocked by another. This blockage has three effects: the amplitudes of all correlations with the affected antenna are decreased, the blocked antenna has an asymmetric primary beam, and, more importantly, the effective interferometer spacing is changed. The amplitude effect is the same as that of the amplitude calibration errors discussed in Section 3.2.2. The error caused by the incorrect effective spacing is like a multiplicative error which increases with distance from the field center (i.e., it is like a scale change). In the image plane, sources will be convolved with an asymmetric error function which increases in amplitude away from the field center. Since a source very near the field center will have almost no error, such a source can not be used to judge the quality of the image further away from the field center. This effect is most important when imaging large fields (small Δu).

3.3.3. Pointing errors Differences in pointing between the elements of the array cause amplitude errors that can be different for each element and can vary with time. Since the magnitude of the error depends on the position of the source in the primary beam, this type of error can not be represented by a convolution and will not be corrected by 'CLEAN' or by the self-calibration techniques unless the region of emission is confined to a small region in the primary beam. The effects of this type of error are strongest near the half-power point of the primary beam, and, since only the amplitude is affected, they will be purely symmetric.

3.4. Computational errors

A number of additional errors can be introduced by the computational methods used to produce a final image. Since these all are discussed in other lectures, I will not repeat the discussion now, but only give a list for completeness. The effects of the approximations used in obtaining the Fourier transform relation and the effects of the aliasing and convolution required to use the fast Fourier transform (FFT) algorithm are discussed in Lecture 7. The effect of having non-coplanar interferometer baselines and finite computing precision are discussed in Lectures 8, 17 and 19. Errors may also be introduced by the image restoration algorithms (e.g., 'CLEAN') and the self-calibration technique; these are discussed in Lectures 8 and 10.

4. Diagnostic Tools

4.1. Low-resolution images

A heavily tapered image covering a large area (the full primary beam) is sufficiently useful that it should be made as the first reduction step. This low-resolution image will give an immediate overview of all the radio emission in the primary beam, and can be used for a number of different purposes:

1. Possible confusing sources which would either alias into a smaller field or have sidelobes in the smaller field (see Lecture 2) can be recognized.

2. Extended emission is more obvious. If unrecognized in a higher-resolution image, it can be mistaken for an error (see Sec. 3.2.1).

3. Various checks and computations (e.g., deconvolution) can be performed quickly, because of the smaller size (in pixels) of the low-resolution image.

4. You may even discover something new and unexpected by looking at the largest possible field of view, and having higher brightness-sensitivity.

4.2. Polarization

Some instrumental errors are highly polarized because they affect only one of the two independent receiver channels. Other errors (e.g., atmosphere, imaging algorithm approximations) and most of the effects of source structure cancel out for all the unpolarized emission. The circular polarization images are especially useful as a diagnostic tool since very little circularly polarized emission is expected for most classes of radio source (see also Lecture 13).

4.3. Fourier transforming the image

In some cases the instrumental errors can be isolated in the image plane. It may be possible either to spatially isolate a region with errors from other sources or to stop the deconvolution, before the errors are reached, and to make use of the residual image. In these cases, the Fourier transform of the errors may show their nature in an obvious way. This technique was used to diagnose the error in the Cas A data shown in Figure 15–2.

4.4. Effective use of image displays

Finally, the effective use of image displays, both in the image and the (u, v) planes, is one of the most useful diagnostic tools available.

5. Acknowledgments

I thank Rick Perley for comments on the text, for discussions, and for extensive help with the Figures.

References

Bracewell, R. N. 1978, *"The Fourier Transform and its Applications"*, Second Edition,McGraw–Hill, New York.

Thompson, A. R. 1982a, *IEEE Trans. Antennas Propagat.*, **AP-30**, 450–456.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

16. Non-Imaging Data Analysis

Timothy J. Pearson
California Institute of Technology, Pasadena, CA 91125, U.S.A.

Abstract. In many types of observation it is impossible or inappropriate to make an image from the visibility data. In this lecture I discuss ways of interpreting visibility data directly, with an emphasis on model-fitting techniques.

1. Introduction

Producing an image with the standard Fourier synthesis and deconvolution procedures is not always the best way to analyze data from a synthesis telescope. In many observations a lot may be learned by inspection of the data in the visibility domain or (u, v) plane. This is, after all, the domain in which the measurements are made, and where errors in the data are easiest to recognize. The Fourier transform involved in imaging spreads errors that are localized in the (u, v) plane throughout the image, so different pixels in an image will have correlated errors, while measurements at different points in the (u, v) plane are largely uncorrelated. Thus quantitative analysis, including estimates of errors in the derived quantities, is often best done in the (u, v) plane. There are some types of observation, particularly those with very sparse (u, v) coverage or poor calibration, in which it is difficult or impossible to make an image, and in these cases it is necessary to interpret the observed visibility data directly. I shall also show that some quantitative astronomical questions can be addressed better in the visibility domain than in the image. For example, in comparing two images made at different times it can be difficult to determine whether apparent changes are due to real changes in the source, or just to differences in the (u, v) plane sampling and the imaging parameters. It is much more straightforward to compare the measured visibilities directly.

In this Lecture I will discuss only continuum data (single frequency channel, single polarization), although the techniques are readily extended to more complex data sets.

2. Visibility Data

As discussed in earlier lectures, the complex visibility $V(u, v)$ measured on a baseline with coordinates (u, v) is related to the Fourier transform of the sky brightness distribution $I(l, m)$:

$$V(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}(l,m) I(l,m) \exp[-2\pi i(ul + vm)]\, dl\, dm. \qquad (16\text{–}1)$$

$\mathcal{A}(l, m)$ is the primary beam response, which can often be ignored when the field of interest is small, as in VLBI. In many cases, it is possible to construct an estimate of the sky brightness distribution by direct inversion of this equation. However, in practice there are many difficulties with this approach which make it preferable to try to interpret the visibility data directly rather than first forming an image.

2.1. Sampling

Measurements of $V(u,v)$ are available at only a finite number of points in a small region of the (u,v) plane. The effect of this is that the reconstructed image is a convolution of the sky brightness with a dirty beam. In extreme cases (such as a single baseline) the dirty beam can have such large and extensive side-lobes that a reliable deconvolution is impossible: too many Fourier components of the image are unconstrained. In such cases an image may be liable to misinterpretation or even impossible to interpret.

2.2. Calibration

The visibility data themselves may be uncalibrated or uncalibratable. For example, the phases of the visibilities may be severely corrupted by the atmosphere or local-oscillator instabilities, or the amplitudes may be dominated by unknown or changing antenna gains. Often this problem can be avoided by using self-calibration, but in extreme cases – again, these cases are often those with sparse (u,v) plane sampling – self-calibration may fail to converge reliably or the result of self-calibration may be dependent on the assumed starting model. In these cases, we can sometimes proceed by examining the *closure phases* and *closure amplitudes*, which are unaffected by the calibration errors. While these quantities are "good observables," they cannot be used to derive an image by simple Fourier transformation, and they can be rather difficult to interpret. Use of the closure quantities is discussed in Section 4.

2.3. Non-Fourier Imaging

In some cases the simple Fourier transform relationship (Eq. 16–1) is not a sufficiently accurate representation of the imaging process; for example in wide-field mapping with non-coplanar baselines where the assumptions used in deriving Eq. 16–1 break down.

2.4. Noise

The observed visibilities are subject to additive noise from the sky, receivers, ground pick-up, etc. To a very good approximation, the noise is Gaussian with equal variance in the real and imaginary parts of the visibility. The Fourier inversion of Eq. 16–1 has the desirable property that the noise in the dirty image is also Gaussian, with known covariance. However, if the image is deconvolved with CLEAN, MEM or similar non-linear techniques, the noise properties of the resulting image will be poorly understood and it may be difficult to estimate the uncertainty in a measurement (e.g., of a component flux density) from the image. For quantitative analysis, it is often better to work directly with the visibilities.

3. Inspecting Visibility Data

All users of synthesis telescopes should at some point look at their visibility data. The visibility data are complex quantities sampled at points spread across the (u,v) plane, so it can be difficult to visualize the whole dataset. Useful displays include plots of amplitude or phase as functions of time on a single baseline, and plots of amplitude and phase versus radius in the (u, v) plane or projected onto a particular direction in the (u, v) plane. Plotting subsets of data involving a single antenna can often identify systematic errors. All of these displays, and more, are readily available in the standard packages (AIPS, AIPS++, DIFMAP). On strong sources, including most that are strong enough for self-calibration, these plots will show structure corresponding to the Fourier transform of the brightness distribution. For weak sources, the structure is usually obscured by the scatter of points due to receiver noise, but suitable averaging of the data (up to the coherence time) may allow source features to be seen.

In order to interpret visibility data, it is essential to have a good understanding of the basic properties of the Fourier transform, which are summarized in the Appendix. With some experience, it is quite easy to recognize features of simple sources by inspection of the visibility data. It is usual to try to represent the source as a sum of simple "components". The *addition theorem* ensures that the visibility function is the sum of the complex visibilities of the individual components. For example, a point source model has a visibility function that has constant amplitude but a phase gradient across the (u, v) plane, the gradient depending on the displacement from the phase center. A double-source model, consisting of two point components, has a visibility function that is the sum of two such functions with different phase gradients; the resultant visibility has a sinusoidal amplitude with wavelength inversely proportional to the double separation. Figure 16–1 is a useful summary that shows how the basic properties of a simple model can be estimated by inspection of the dependence of visibility amplitude and phase on position in the (u, v) plane.

Several simple "component" brightness distributions are frequently used (see Appendix). The most common is the Gaussian, which has a simple Fourier transform – another Gaussian. The uniform disk and the optically-thin sphere are physically somewhat more plausible models, whose Fourier transforms can be expressed in terms of Bessel functions. However, if the components are only barely resolved, the exact functional form of the brightness distribution is unimportant: all these functions have a similar quadratic dependence on baseline at short baselines. Figure 16–2 shows the dependence of visibility amplitude on baseline length for several brightness distributions, adjusted so that the 50%-visibility points coincide. This figure shows, for example, that at short baselines a Gaussian of FWHM 1 arcsec is indistinguishable from an optically thin sphere of diameter 1.8 arcsec.

In order to compare a model with an image derived, for example, by Fourier inversion and self-calibration, the model should be convolved with a point-spread function (or "beam"). As in CLEAN, it is conventional to use a Gaussian beam. If the model is made up of Gaussian components, this is straight forward: the convolution of two elliptical Gaussians is another elliptical Gaussian whose parameters can be determined analytically (Wild 1970). (For circular Gaussians, the widths just add quadratically.) For the other component types, a numerical convolution must be used.

3.1. An Example

As an example, I will consider a data set from a VLBI observation in which the source 2021+614 was observed in four hour-long "snapshots" on 11 antennas.

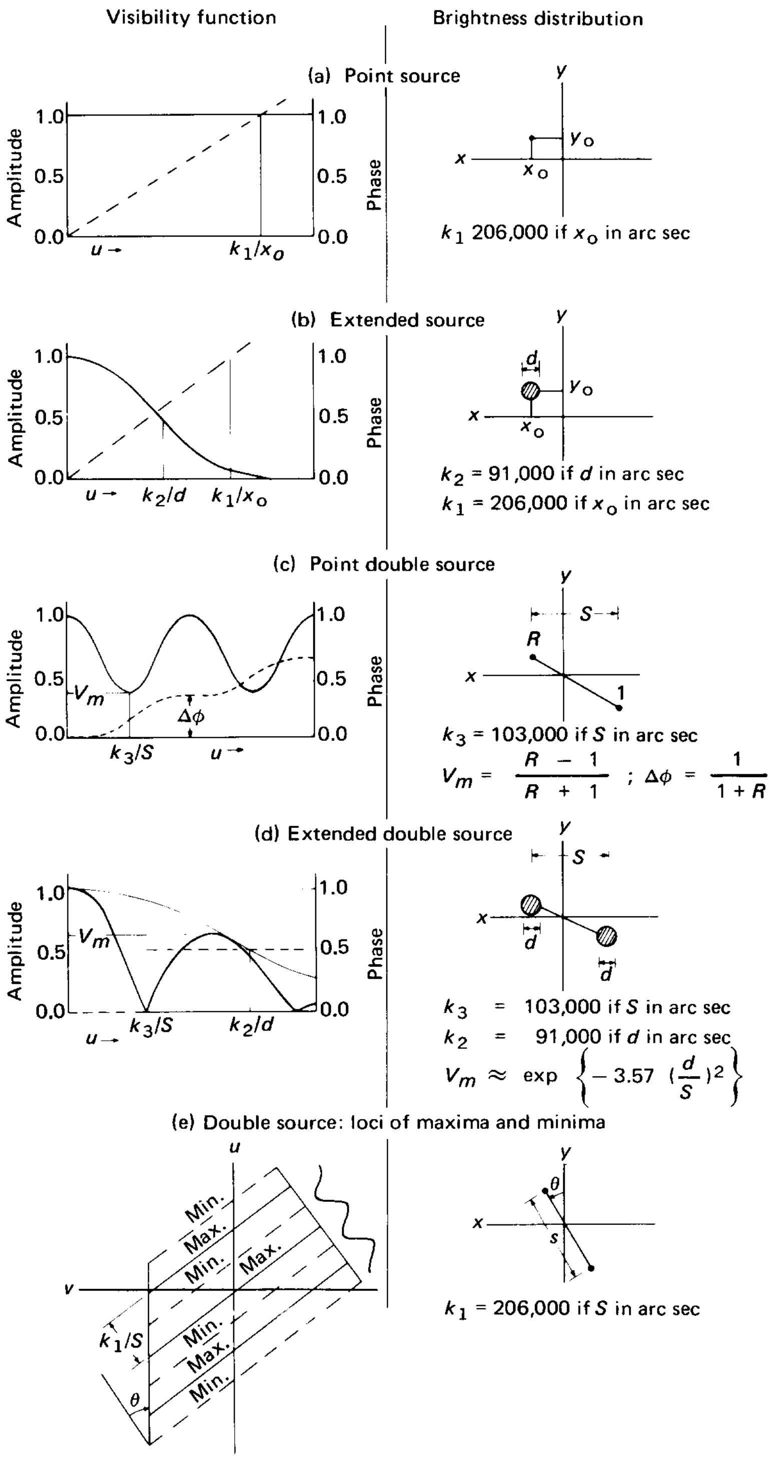

Visibility function
Brightness distribution
(a) Point source
Amplitude
Phase
1.0
0.5
0.0
u →
k_1/x_o
y
x
y_o
x_o
k_1 206,000 if x_o in arc sec
(b) Extended source
k_2/d
k_1/x_o
d
k_2 = 91,000 if d in arc sec
k_1 = 206,000 if x_o in arc sec
(c) Point double source
V_m
Δφ
k_3/S
S
R
1
k_3 = 103,000 if S in arc sec
$V_m = \frac{R-1}{R+1}$; $\Delta\phi = \frac{1}{1+R}$
(d) Extended double source
k_3 = 103,000 if S in arc sec
k_2 = 91,000 if d in arc sec
$V_m \approx \exp\left\{-3.57\left(\frac{d}{S}\right)^2\right\}$
(e) Double source: loci of maxima and minima
u
v
Min.
Max.
k_1/S
θ
s
k_1 = 206,000 if S in arc sec

Figure 16–1. The visibility functions for various brightness distribution models (reproduced from Fomalont & Wright 1974; © Springer, Berlin).

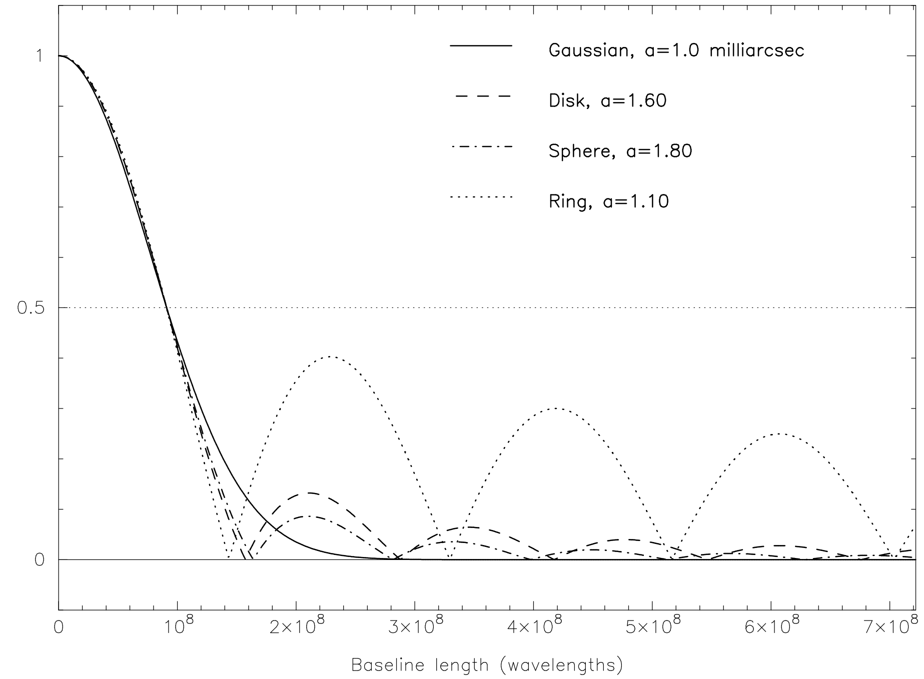


Figure 16–2. Dependence of visibility amplitude on baseline length for four different circularly symmetric brightness distributions. Analytical expressions for these functions are given in the Appendix. The scale-size a for each distribution has been adjusted so that the visibility of each drops to 50% at the same baseline.

There are many ways of examining the visibility data from such an observation, and I encourage you to use all the tools at your disposal to do so. Some of the most useful projections of the two-dimensional complex data set are shown in Figure 16–3, which was obtained with the program DIFMAP (Shepherd 1997). A simple plot of the (u, v) coverage (Fig. 16–3a) shows the spatial frequencies to which the observation is sensitive. It can also be useful to encode visibility amplitude information by color or symbol size. The graph of amplitude versus projected baseline length (Fig. 16–3b) in this case shows that the source is resolved (not dominated by a point component of constant amplitude) and has subcomponents that beat against each other to give a wide range of visibility amplitudes at each baseline. By projecting onto a line in the (u, v) plane (Fig. 16–3c) we can see that to first approximation the source is an equal double; after adjusting the projection angle to make the minima line up, we find that the visibility takes on the canonical form shown in Fig. 16–1d. By comparison with Fig. 16–1d, we find an approximate *starting model*: two equal components, each of 1.25 Jy, separated by about 6.8 mas in p.a. 33°. From the upper envelope, the size of each component is about 0.8 mas (Gaussian FWHM).

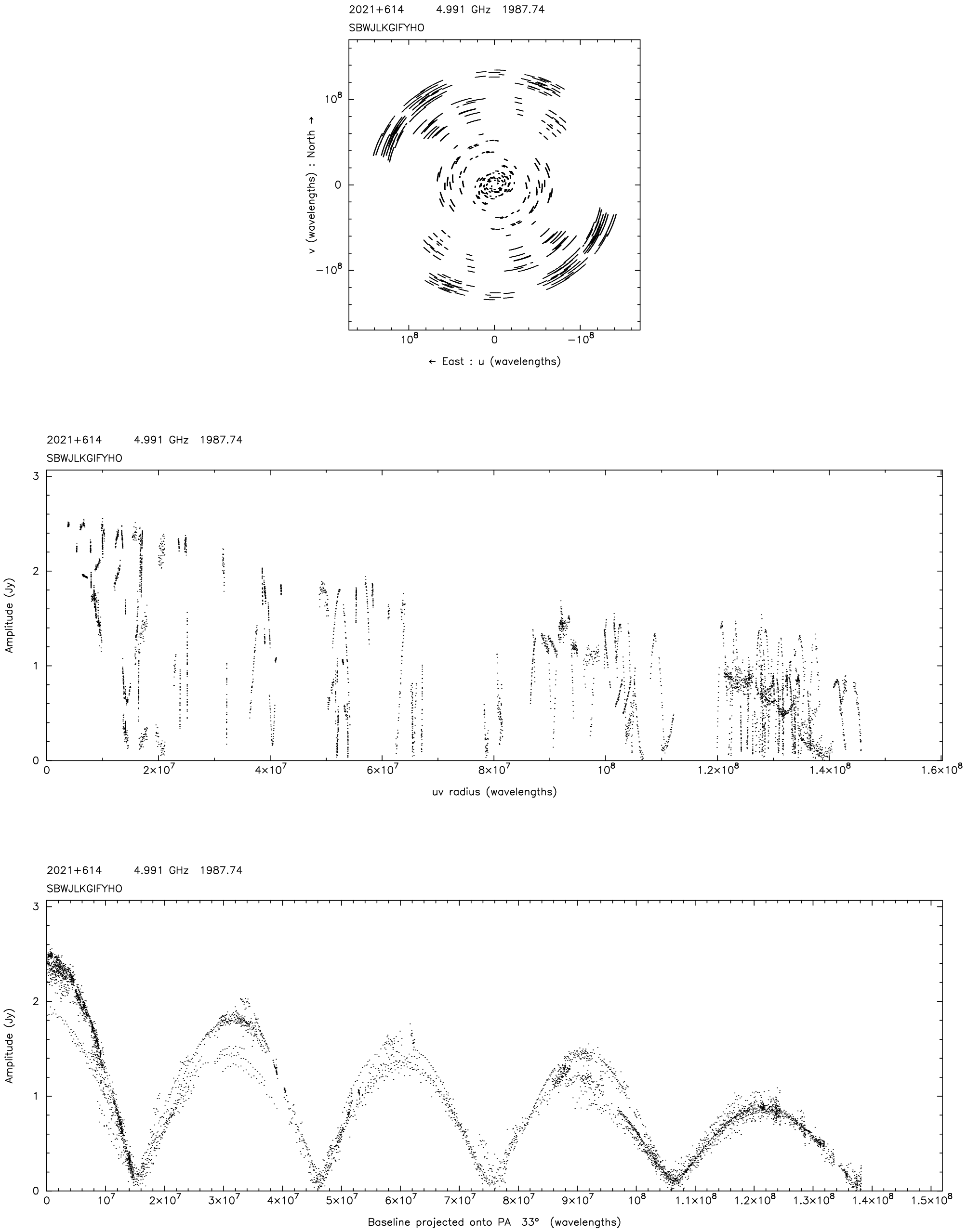


Figure 16–3. Observed visibility data from a 5-GHz Mark-II VLBI observation of the radio galaxy 2021+614 in 1987 (Conway et al. 1994). (a) The sampling of the u, v plane. (b) Visibility amplitude as a function of radius in the u, v plane (projected baseline). (c) Amplitude as a function of the component of baseline length projected in position angle 33°.

4. The Closure Quantities

The closure phase (Jennison 1958; Rogers et al. 1974) is the sum of the visibility phases around a triangle of three baselines:

$$\Psi_{lmn}(t) = \phi_{lm}(t) + \phi_{mn}(t) + \phi_{nl}(t), \tag{16–2}$$

where $\phi_{lm}(t)$ is the visibility phase on the baseline between antennas l and m at time t. In this sum, antenna-based phase errors cancel (Pearson & Readhead 1984). The closure phase is the argument of the *bispectrum* of the sky brightness distribution, which is the triple product of complex visibilities on a closed triangle of baselines:

$$\text{Bispectrum} = V(u,v)V(u',v')V(-u-u',-v-v'). \tag{16–3}$$

This makes it clear that the closure phase is a function of four variables (two positions in the (u,v) plane). In practice this four-dimensional space is always poorly sampled, and the relationship of the observed closure phases to the sky brightness distribution is far from intuitive. However, in an array of N antennas with $N(N-1)/2$ baselines, the fraction of the visibility phase information that is available from the closure phases is $(N-2)/N$.

Another "good observable" is the closure amplitude (Twiss, Carter, & Little 1960; Readhead et al. 1980). This is the ratio of visibility amplitudes on baselines between four antennas arranged so that antenna gains cancel:

$$\frac{|V_{kl}| \cdot |V_{mn}|}{|V_{km}| \cdot |V_{ln}|}. \tag{16–4}$$

The closure amplitude is a function of six variables (three points in the (u,v) plane). The fraction of the amplitude information available from the closure amplitudes is $(N-3)/(N-1)$.

In many cases it is possible to construct an image from the closure quantities by the iterative self-calibration procedures discussed in other lectures, but if the data are sparse it is often better to interpret the closure quantities directly using the methods outlined in this lecture.

Several programs are available for inspection of closure phases and closure amplitudes, but because they are functions of 4 or 6 variables, it is rarely possible to learn much about the structure of the source simply by inspection of the closure quantities. An exception is a double source which will show large jumps in the closure phase where the amplitude on one baseline has a minimum.

5. Model Fitting

5.1. Imaging as an Inverse Process

Synthesis imaging is a member of the general class of *inverse problems*. In such problems, we understand the *forward problem*, which in our case means that if we knew the true sky brightness we could calculate the measured quantities (complex visibilities or closure quantities) using Eq. 16–1 or a suitable generalization, but we want to invert this process to estimate the sky brightness from the measurements. There is a wide literature on inverse problems which emphasizes the difficulties of determining whether there is a unique solution and of devising a stable algorithm that will find the solution (e.g., Parker 1977; Press et al. 1992).

One technique that is generally applicable to inverse problems is *model fitting.* In this technique, one makes a parametric model of the sky brightness distribution and uses the imaging equations to calculate the expected measurements. One then adjusts the parameters of the model to get the "best fit" model. When the error statistics of the observations are understood, the appropriate statistical technique to estimate the parameters of the model is the *maximum likelihood* method. For Gaussian errors, this is the same as the *least-squares* method. Press et al. (1992, chapter 15, or chapter 14 in the first edition) give a more extended discussion of least-squares fitting that I strongly recommend to anyone who uses a least-squares program; it is of course essential reading for anyone planning to write such a program. Another useful reference is the book by Bevington & Robinson (1992, chapter 8); the first edition of this book (Bevington 1969, chapter 11) is still useful for Fortran programmers.

There are three steps involved in model fitting: (1) Design a *model* defined by a number of adjustable parameters; (2) Choose a *figure-of-merit* function; (3) Adjust the parameters to *minimize the merit function.* The goals are to obtain: (1) Best-fit values for the parameters; (2) A measure of the goodness-of-fit of the optimized model (relative to the measurement errors); (3) Estimates of the uncertainty in the best-fit parameters.

As an example, consider a model of the N observed visibilities $V_i(u,v)$ (the same technique is applicable to the closure quantities). The model, $F(u,v)$, depends on a number M of parameters a_j (typically there will be 6 parameters per model "component": component flux density, two sky coordinates, angular size, axial ratio, and orientation). The model is intended to reproduce the observations within their uncertainty, i.e.,

$$V(u,v) = F(u,v;a_1,\ldots,a_M) + \text{noise}. \tag{16–5}$$

The *likelihood* of the model is the probability of obtaining the data, assuming that the model is correct. If the noise in the observations is Gaussian, so that each visibility measurement V_i has an associated standard deviation σ_i, the likelihood is:

$$L \propto \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{V_i - F(u_i,v_i;a_1,\ldots,a_M)}{\sigma_i} \right)^2 \right] \right\}. \tag{16–6}$$

The conventional method of statistical estimation is the *maximum likelihood method*: choose the values of the parameters that maximize L. This is equivalent to minimizing $-\log L$, or minimizing

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{V_i - F(u_i,v_i;a_1,\ldots,a_M)}{\sigma_i} \right)^2 \tag{16–7}$$

Thus the maximum likelihood method is, in this case, the method of least squares: we must minimize χ^2, the weighted sum of squares of the deviations between the data and the model. Note that this is strictly applicable only if the data have Gaussian errors.

If the errors are Gaussian, then near the minimum χ^2 follows the chi-square distribution with $\nu = N - M$ degrees of freedom (N data points, M parameters). The expected value of χ^2 is ν with standard deviation $\sqrt{2\nu}$. A measure of the goodness of fit of the optimized model is the *reduced chi-square* $\chi^2/(N - M)$ which should be close to 1 for a good fit. Large values of the reduced chi-square indicate a bad fit, while values much smaller than 1 indicate too good a fit, perhaps because the errors σ_i have been overestimated.

Note that the measured quantities V_i need not be the complex visibilities (or amplitudes and phases): one might for example choose to use the closure quantities if the data set is poorly calibrated.

Model fitting has many desirable properties. It can (in principle) take into account all the details of the measurement process, which need not be a simple Fourier transform; and because it operates in the domain of observation, where the statistics of the measurement process are well understood, it can estimate the statistical uncertainties in the parameters of the best-fit model, which may be the astronomically important quantities. However, it also has serious problems: it may be difficult to choose a suitable parameterization of the model, the solutions are not unique, and it can be much, much slower than the conventional Fourier inversion and deconvolution methods.

5.2. Uses of Model Fitting

Fitting a model to the visibility data is not the appropriate technique to use for every observation, of course. If the primary objective is an image, and there are sufficient data to form a reliable image, then use the standard inversion and deconvolution techniques. Model fitting is very useful, however, for interpreting sparse or uncalibrated data, and for quantitative analysis. In many cases model fitting and conventional imaging will both be useful and can supplement each other. Model fitting is most useful, of course, when the brightness distribution can be represented accurately by a model with a small number of parameters. Examples include fields containing a small number of unresolved sources, where the parameters are the positions and flux densities of the sources; barely resolved sources, where the parameters are position, flux density, and angular size (if the source is not more than 50% resolved on the longest baseline, three parameters – major axis, minor axis, and position angle – are sufficient to characterize the brightness distribution: see Figure 16–2); and sources consisting of a small number of barely-resolved components.

Model fitting is typically used for the following:

1. Checking and adjusting amplitude calibration. Good unresolved calibration sources are difficult to find, especially for VLBI observations, and frequently a calibrator will have some structure. However in many cases the calibrator can be represented accurately by a one- or two-component model. The parameters of such a model can be estimated by model fitting using the best-calibrated baselines of the array, and then the model can be used to calibrate the other antennas. This technique has proved particularly useful on arrays including antennas with a wide range of sensitivities. It is advisable to check the calibration by using two or more calibrators.

2. The hybrid-mapping and self-calibration methods for making images from poorly calibrated or uncalibrated data all require a starting model. Frequently a point-source model can be used, but I have found that the self-calibration process always converges faster if a good starting model is used. This is particularly important for sources that are very dissimilar from a point source, such as almost-equal double sources. A good starting model is also advisable for the Schwab–Cotton global fringe-fitting algorithm used in AIPS task FRING (Schwab & Cotton 1983); indeed if the source is not close to a point source, this algorithm may fail to find fringes altogether if a point-source model is used. One technique that I have found effective is to make a crude image by the standard self-calibration technique starting from a point source, and use the image to deduce a starting model for least-squares model fitting, the result of which is used in turn as the starting point for another round of self-calibration (Biretta, Moore, & Cohen 1986). Use of model fitting in this way can be regarded as the application of another constraint in the self-calibration process; in addition to enforcing positivity and finite support, we are requiring the image to be "simple" and "smooth." Using model fitting instead of CLEAN to refine the self-calibration model thus shares some of the advantages of the Non-Negative Least Squares algorithm (Briggs 1995).

3. If the source is simple enough to be accurately represented by a parametric model, then model fitting provides accurate estimates of the model parameters, such as component flux densities and separations, with error estimates. Estimates derived in this way directly from the visibility data will almost always be more reliable than parameters estimated from a deconvolved image. Model fitting is also a deconvolution process: the component size estimates or positional uncertainties may be much smaller than the beam, but they can still be reliable if the signal-to-noise ratio is high enough. The error estimates should provide a guide to the reliability of the deconvolution. The drawback of this approach is that the form of the chosen model may not be correct, and almost certainly will not be unique; it is necessary to explore a range of different model types.

4. Model fitting can be used to improve the behavior of the commonly used imaging techniques CLEAN and MEM. For example, for an extended source CLEAN often gives better results if a large part of the extended structure can be modeled and subtracted before deconvolution; MEM can give better results if dominant point sources are modeled and subtracted.

5. The limiting factor in astrometric and geodetic VLBI is often resolved structure in the reference sources. Rather than try to image each source from the astrometric data themselves (which are usually not planned for optimum imaging), it may be better to use a simple parametric model of each source adjusted to fit the complete ensemble of observations; if necessary, the parameters can be adjusted to take into account smoothly varying source structure.

5.3. Practical Model Fitting

I now return to the example dataset introduced in Section 3.1. The simple model that we derived by visual inspection is shown in Table 1. We can now use a least-squares model-fitting program to improve this model. This data set has been calibrated in amplitude but not in phase, so we fit to the visibility amplitudes and closure phases. The initial "eyeball" model has agreement factors (square root of reduced chi-square) of 5.9 (amplitudes) and 4.2 (closure phases). After adjusting the parameters the agreement factors are improved to 3.9 and 3.1. Further improvement requires the introduction of more components; after some perseverance, I was able to obtain the 5-component model shown in Table 16–1 and Figure 16–4, with agreement factors of 1.3 and 1.5.

Table 16–1. Model Parameters

	Flux	Radius	Theta	Axis	Ratio	Phi
Starting ("eyeball") model						
	1.250	0.000	0.0	0.80	1.00	0.0
	1.250	6.800	33.0	0.80	1.00	0.0
Best 2-component model						
	1.185	0.000	0.0	0.81	0.76	45.1
	1.141	6.788	32.9	0.89	0.79	54.0
Best 5-component model						
	1.008	0.000	0.0	0.70	0.73	46.7
	1.094	6.786	32.9	0.89	0.75	56.1
	0.142	1.169	45.2	1.79	0.25	-17.7
	0.128	9.378	41.4	0.78	0.77	19.0
	0.120	1.893	72.6	3.89	0.00	55.3

The parameters given for each component are: flux density (Jy), position of center (polar coordinates relative to phase center, with radius in milliarcsec), FWHM major axis (milliarcsec), axial ratio (minor/major), and position angle of major axis.

5.4. Programs for Model Fitting

Several programs are available for model fitting using the complex visibilities. In the AIPS package, the task UVFIT does a least-squares fit of a model to complex visibility data (real part and imaginary part); this is preferable to fitting to amplitude and phase for reasons discussed below. SLIME (Flatters 1998) is an add-on AIPS task that combines model fitting with a graphical editor. Martin Shepherd's program DIFMAP also provides a convenient graphical interface for model fitting and data inspection. All of these programs allow the model to be defined as a sum of Gaussian, optically-thin sphere, or other components (see the Appendix). If the data are poorly calibrated, model fitting and self-calibration must be alternated in an iterative process, in order to derive both the model and the calibration parameters.

In some cases, a better approach to uncalibrated data would be to fit to the closure quantities directly, or to amplitudes and closure phases if only the phases are badly calibrated. Because there can be a very large number of closure

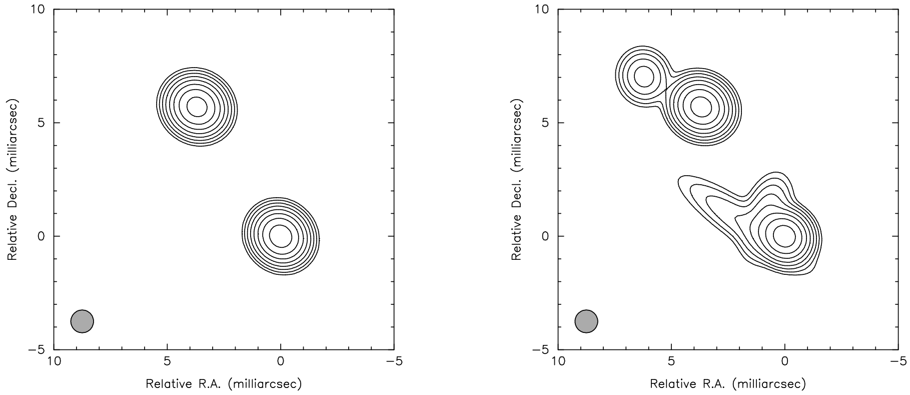


Figure 16–4. Contour maps of two Gaussian models of 2021+614, convolved with a circular Gaussian beam of FWHM 1 milliarcsec. Contours are drawn at 0.5, 1, 2, 4, . . ., 64% of the peak. *Left*: best two-component model; *right*: best five-component model.

quantities (although they are not all independent), this is less well supported by existing programs. One program, MODELFIT, originally written by George Purcell and Richard Simon, is available in the Caltech VLBI package (Pearson 1991); however, this program is now showing its age and cannot cope with modern multichannel datasets.

Similar least-squares model fitting can be applied in the image plane. The AIPS task JMFIT can fit one or more Gaussian components to a region of an image. This can sometimes be a useful way to derive a starting model for (u, v) plane fitting. Image-plane fitting is most useful for estimating the parameters of a source when there is complex emission elsewhere in the field. However, a proper error analysis, taking into account correlations between pixels, is difficult.

5.5. Limitations of the Least-Squares Method

Note that when we use the least-squares method, we are making the following assumptions:

1. The model is actually a good fit to the data. This should be tested by checking that the reduced chi-square is close to 1.

2. The errors are actually Gaussian. This is true if the data are the real and imaginary parts of the observed visibility, but not if the data are visibility amplitudes, closure phases, or closure amplitudes, which do not have Gaussian distributions except in the limit of high signal-to-noise ratio. The least-squares method is frequently (incorrectly) used with non-Gaussian data, but it must be used with considerable caution: the results may be biased and any estimates of the errors on the fitted parameters may be wrong.

3. The errors are known. In most cases, an accurate estimate of the errors in the real and imaginary parts of the visibilities should be available from the system temperatures, from the statistics of bit counting in the correlator, or from the scatter of points within an integration. However, this information is often discarded during data reduction and may be difficult to recover.

4. There are no systematic (calibration) errors. Systematic errors that are not removed in the calibration can in principle be estimated as additional parameters in the model; e.g., one could regard the antenna gain factors as unknown parameters to be estimated. When data have been self-calibrated in amplitude or phase before model fitting, additional parameters (antenna gains or phases) have been estimated from the data. This reduces the number of degrees of freedom. If this is not taken into account when calculating the reduced chi-square, then the model will appear to fit the data better than it should (reduced chi-square too small).

5. The errors are uncorrelated. This should be the case for additive noise in the observed visibilities, but it is not true, for example, of errors in the closure phases or closure amplitudes if non-independent closure quantities are included in the fit.

5.6. Least-Squares Algorithms

The goal of any least-squares algorithm is to find the global minimum of χ^2 in the M-dimensional parameter space. At any minimum, the derivatives of χ^2 with respect to the parameters will be zero,

$$\nabla\chi^2 = \frac{\partial\chi^2}{\partial a_k} = 0. \tag{16–8}$$

If the model $F(u, v; a_i, \ldots, a_M)$ is a linear function of the parameters a_i, this set of M equations can be solved by standard matrix-inversion methods. In most cases, however, the model is a non-linear function of the parameters and an iterative technique must be used.

One simple method is the "grid search" (Bevington & Robinson 1992): select starting trial values for the parameters, and adjust each parameter in turn (keeping the others fixed) to minimize χ^2 with respect to that parameter. The disadvantages of this method are that it can be very slow, and it is not obvious by how much one should change each parameter at each trial.

A more efficient method is the "gradient search": at each step increment all the parameters to move in the direction of the gradient of χ^2 in parameter space. The gradient of χ^2 is usually calculated numerically by evaluating the change in χ^2 for small increments of each parameter; thus one step in the gradient search may require many more model evaluations than one step in the grid search, but it should take one closer to the minimum. Alternatively, $\nabla\chi^2$ can be determined directly if analytical expressions for the derivative of the model with respect to each parameter are available (DIFMAP uses this method: all the standard model component shapes have analytical derivatives). The magnitude of the step to take along the gradient can be estimated by examining the second derivative of χ^2, the *Hessian* matrix

$$\nabla^2 \chi^2 = \frac{\partial^2 \chi^2}{\partial a_k \partial a_l} \tag{16–9}$$

The most commonly used algorithm is a refinement of this simple gradient search known as the Levenberg-Marquardt method; for details, see Press et al. (1992) and Bevington & Robinson (1992). More sophisticated algorithms are also available (e.g., Bunch, Gay, & Welsch 1993).

In practice, you are likely to run into a number of problems whatever algorithm you use:

1. Finding the global minimum. It is very easy to get stuck in a local minimum of χ^2 which may be far from the global minimum (where the reduced chi-square should be close to 1). The better the starting point, the more likely you are to find the global minimum, which is why you should spend some time refining the starting model by inspection of the visibilities before using a least-squares program. The grid search method can sometimes find its way out of a local minimum where the gradient search method gets stuck.

2. Slow convergence. Often the minimum χ^2 lies in a wide flat valley, where changes to some of the parameters make little change to χ^2. Gradient search should converge faster than grid search in this case. However, this is often a symptom of a poorly-constrained model where parameters are not independent. For example, if data are available in only a limited range of (u, v) distance, a strong, wide component may be difficult to distinguish from a weaker, more compact component. In such cases it is best to constrain some of the parameters to their *a priori* values and let the program adjust the others; but remember that the result is not a unique solution.

3. Constraints. A physically realistic model will have positive component flux densities, and the model-fitting program should apply such constraints. Other constraints may be artificial ones imposed by the formulation of the model; e.g., if the component position is specified by polar coordinates (r, θ), r should be positive, and θ is poorly constrained when r is small. A better approach might be to choose orthogonal coordinates $(x, y) = (r\cos\theta, r\sin\theta)$ as the adjustable parameters. Similar considerations apply to major axis, axial ratio, and component position angle: the algorithm often finds its best solution at a boundary of parameter space with a zero axial ratio (this is apparent in the model in Table 1).

4. Choosing the right number of parameters or components. Model fitting does not have a unique solution, and with a sufficient number of variable parameters many equally good solutions can be found. One should not introduce more parameters than are necessary to obtain a reduced chi-square close to unity. Thus if a circular Gaussian component is a good fit, it is not necessary to adjust the axial ratio and position angle. The appropriate statistical test for determining whether additional parameters

actually improve the fit is the F-test (Bevington & Robinson 1992, chapter 11). However, such tests are not very useful if the assumption that the data points have independent Gaussian noise of known magnitude is not valid.

5.7. Error Estimation

Press et al. (1992, section 15.6) give a good account of methods for determining confidence limits on the estimated model parameters. Most model-fitting programs determine the curvature of the χ^2 surface around the minimum, from which a covariance matrix for the fitted parameters can be determined. It is a good idea to look at the covariance matrix to see which parameters are well constrained and how the parameters are correlated. However, it is dangerous to use the covariance matrix for direct estimates of the uncertainties of the parameters, because the assumptions that go into the theory of least-squares are frequently violated. A better approach is to use contours of constant chi-square around the minimum to define confidence limits on the parameters, i.e., find the region of parameter space in which

$$\chi^2 < \chi^2_{\rm min} + \Delta\chi^2. \tag{16–10}$$

The choice of $\Delta\chi^2$ depends on the required *confidence level* and the number of parameters estimated; e.g., for 68.3% confidence and one parameter, $\Delta\chi^2 = 1$; for 90% confidence and 6 parameters, $\Delta\chi^2 = 16.8$. One must distinguish confidence intervals on a set of parameters considered jointly from confidence intervals on a single parameter. In the latter case, the goal is to determine how bad the fit gets as the specified parameter is changed from its optimum value, while adjusting the other parameters to compensate. The 68.3% confidence range for a single parameter can be found by projecting the contour $\Delta\chi^2 = 1$ onto the axis corresponding to that parameter.

It is important to remember that these theoretical confidence limits will not apply if the data are not Gaussian or not independent, or if the best-fit model is not in fact a good fit. It is a good idea to test these assumptions by other methods. The Monte-Carlo method involves using a model of the sky brightness distribution to generate simulated data sets observed under similar conditions to the real observations. By applying the model-fitting procedure to these data sets, and comparing the results with the input model, one can get some idea of the uncertainties involved. The uncertainties are likely to be larger for real data sets which may contain errors of unknown origin that cannot be simulated. Another approach is to obtain multiple data sets by repeated observation, or by dividing a data set into subsets.

6. Some Applications

6.1. Superluminal Motion

An important application of visibility analysis is in the detection and measurement of *changes* in the brightness distribution of a source. For example, in superluminal sources we compare the relative positions of components at different epochs in order to measure an apparent expansion speed. Indeed, superluminal motion was first discovered, in the quasar 3C 279, in VLBI observations on a single baseline: sharp minima in the visibility were seen to move toward the origin in the (u, v) plane, corresponding to an expansion of a two-component model of the brightness distribution (Whitney et al. 1971).

Even with good (u, v) coverage, using a direct comparison of images made with the standard self-calibration and deconvolution procedures it can be difficult to disentangle real changes in the source from differences between the images due to calibration errors or differences in (u, v) coverage. A better approach is to compare the visibilities or closure quantities directly, using model fitting as a technique for overcoming the differences in (u, v) coverage. First find a model that is a good fit to the data obtained in one observation (A, say). Usually this will involve imaging and self-calibrating the data set and using the image as a guide to choosing the model. The model can then be transformed to the (u, v) plane and compared directly with the measured visibility samples from the other observation (B). Of course this technique should only be used to interpolate between small differences in (u, v) coverage, not to extrapolate to very different regions of the (u, v) plane. Calibration errors can be circumvented by comparing the closure quantities: significant changes in closure phases or closure amplitudes should be a reliable indicator of real changes in the source. If differences between the two observations are detected in this way, the model can be used to suggest a physical interpretation of the changes between A and B: for example, which components have changed, and are the changes in component flux density, size, or location?

Conway et al. (1994) used this technique to compare observations of the radio galaxy 2021+614 made in 1982 and 1987 (the data presented in Figure 16–3). Although there are significant differences in visibility between the two data sets on the few baselines in common, images made independently from the two data sets are very similar; but by a careful visibility analysis using model fitting, Conway et al. detected an expansion of 69 ± 12 microarcsec, corresponding to a sub-luminal expansion at $0.13c$ (assuming a Hubble constant of 100 km s^{-1} Mpc^{-1}). Note that the estimated uncertainty (due to thermal noise) is much smaller than the beam size (about 1 milliarcsec); this is due to the high signal-to-noise ratio of the two bright components. The greatest remaining uncertainty is the effect of calibration errors.

A similar technique has been used to measure the expansion speeds and distances of planetary nebulae using the VLA (Masson 1986; Kawamura & Masson 1996).

6.2. Gravitational Lenses

A gravitational lens can split the image of a background source into two or more components. If the background source varies in intensity, there will be a time delay between the corresponding variations in the individual image components. Together with the redshifts of the lens and background source, and a model of the gravitational potential of the lens, the time delay can be used to determine the linear scale of the lens, and hence its distance by comparison with the angular scale. This is a powerful technique for measuring the Hubble constant. Many of the best lensed systems are radio sources and the images can be monitored with the VLA or MERLIN to detect variations. It is necessary to measure the component flux densities with high accuracy ($< 1\%$) over a long period, although it may be impossible to use the same (u, v) coverage for every observation. This is an ideal case for model fitting: there are few free parameters (flux densities of a few point sources of known position, but not well resolved from each other), and we need reliable error estimates. In some cases it may be necessary to include unrelated field sources in the model. This technique has been used, for example, to estimate intercomponent time delays in the lens system B1608+656 (Myers et al. 1995; Fassnacht et al., in preparation).

6.3. The Sunyaev-Zeldovich Effect

In the Sunyaev-Zeldovich effect, inverse Compton scattering of photons of the cosmic microwave background by hot electrons in clusters of galaxies shifts the photons to higher energies. The result is a decrement in the background intensity at centimeter wavelengths (e.g., Birkinshaw 1998). The magnitude of the decrement is proportional to the integral of the electron density through the cluster. The electron density can be estimated from X-ray observations, so this gives a measure of the linear depth of the cluster. Comparison with the angular size of the cluster then gives an estimate of the distance and the Hubble constant. Several groups have made synthesis observations of clusters of galaxies to measure the decrement (e.g., Carlstrom, Joy, & Grego 1996 used the Owens Valley millimeter array; Jones et al. 1993 used the Ryle Telescope at Cambridge). The typical angular scale is a few arc minutes, so short baselines are needed.

Typically clusters are modeled with an isothermal β model, for which the expected profile in the decrement is

$$f(r) = f_0 \left(1 + \frac{r^2}{a^2}\right)^{(1-3\beta)/2}, \tag{16–11}$$

where r is the angle from the cluster center and a is a core radius. The visibility, given by a Hankel transform (see the Appendix), is

$$F(\rho) = (2\pi f_0 a^2) \frac{(\pi a \rho)^{n-1}}{\Gamma(n)} K_{1-n}(2\pi a \rho), \tag{16–12}$$

where $n = (3\beta - 1)/2$ and K is the modified Bessel function of the second kind. The visibility decreases exponentially with baseline; e.g., for the typical case $\beta = 2/3$,

$$F(\rho) = (2\pi f_0 a^2) \frac{\exp(-2\pi a \rho)}{2\pi a \rho}. \tag{16–13}$$

The steep increase in visibility to short baselines, and the central "hole" in the (u, v) coverage of all synthesis arrays combine to ensure that a synthesis image will almost always underestimate the magnitude of the decrement. Quantitative analysis requires model fitting to provide estimates of the central decrement and β.

6.4. The Cosmic Microwave Background Radiation

Several groups are conducting or planning interferometric observations of the fluctuations in the microwave background radiation (e.g., the Cambridge Cosmic Anisotropy Telescope: Scott et al. 1996). Although it is possible to make images, this is another field in which the analysis is best done in the (u, v) plane. The background radiation is expected to show Gaussian fluctuations with characteristic angular scales from a few arc minutes to a few degrees. The angular power spectrum of the fluctuations depends sensitively on the cosmological parameters (including the density parameters Ω, Ω_b, and Ω_Λ and the Hubble constant) and an accurate measurement of the power spectrum would be of great value in measuring these parameters. The power spectrum is the square of the Fourier transform of the sky brightness distribution, and is thus closely related to visibility: the expected value of the square of the visibility at (u, v) is proportional to the power spectrum $C(u, v)$ convolved with the square of the Fourier transform of the primary beam $\mathcal{A}(l, m)$. (Here $C(u, v)$ corresponds to the commonly used spherical power spectrum C_l at multipole order $l \approx 2\pi\sqrt{u^2 + v^2}$). Optimum methods to extract the power spectrum from visibility measurements in the presence of noise and possible foreground emission are being developed by several groups (e.g., Hobson, Lasenby, & Jones 1995; Maisinger, Hobson, & Lasenby 1997; White, Carlstrom, & Dragovan 1998).

7. Appendix

7.1. Properties of the Fourier Transform

Given a model brightness distribution $f(l, m)$ in the sky plane, the model visibility $F(u, v)$ is computed by Fourier transformation:

$$F(u, v) = \mathfrak{F}\{f(l, m)\} \tag{16–14}$$

i.e.,

$$F(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(l, m) \exp[-2\pi i(ul + vm)]\, dl\, dm \tag{16–15}$$

The following properties of the Fourier transform are useful for interpreting visibility data. For details, see Bracewell (1986).

Addition Theorem

$$\mathfrak{F}\{f(l, m) + g(l, m)\} = F(u, v) + G(u, v) \tag{16–16}$$

Convolution

$$\mathfrak{F}\{f(l, m) \star g(l, m)\} = F(u, v) \cdot G(u, v) \tag{16–17}$$

Shift Theorem

$$\mathfrak{F}\{f(l - l_i, m - m_i)\} = F(u, v)\ \exp[-2\pi i(ul_i + vm_i)] \tag{16–18}$$

Similarity Theorem

$$\mathfrak{F}\{f(al, bm)\} = \frac{1}{|ab|} F\left(\frac{u}{a}, \frac{v}{b}\right) \tag{16–19}$$

7.2. Simple Models

Here I give analytic expressions for some of the commonly-used model components (Purcell 1973). All the expressions are for a circularly symmetric component $f(r)$ centered at the origin of the (l,m) coordinate system, with $r = \sqrt{l^2+m^2}$. The Fourier transform $F(\rho)$ is circularly symmetric in the (u,v) plane, with $\rho = \sqrt{u^2+v^2}$. The relationship between $f(r)$ and $F(\rho)$ is a Hankel transform:

$$F(\rho) = 2\pi \int_0^\infty f(r) J_0(2\pi r\rho) r\, dr \qquad (16\text{–}20)$$

By application of the theorems of Section 7.1., it is straightforward to derive the expressions for elliptical components with arbitrary positions and orientations.

Delta Function (Point Source)

$$\begin{aligned} f(x,y) &= \delta(x,y), && (16\text{–}21) \\ F(u,v) &= 1. && (16\text{–}22) \end{aligned}$$

Gaussian

$$\begin{aligned} f(r) &= \frac{1}{\sqrt{\pi/4\ln 2a}} \exp\left(\frac{-4\ln 2\, r^2}{a^2}\right) && (16\text{–}23) \\ F(\rho) &= \exp\left(\frac{-(\pi a\rho)^2}{4\ln 2}\right), && (16\text{–}24) \end{aligned}$$

where a = full width to half-maximum intensity (FWHM).

Uniformly Bright Disk

$$\begin{aligned} f(r) &= \begin{cases} 4/(\pi a^2), & \text{if } r \le a/2 \\ 0, & \text{otherwise} \end{cases} && (16\text{–}25) \\ F(\rho) &= \frac{2J_1(\pi a\rho)}{\pi a\rho}, && (16\text{–}26) \end{aligned}$$

where a = diameter.

Optically Thin Sphere (The brightness at each point is proportional to the path length through the sphere.)

$$\begin{aligned} f(r) &= \begin{cases} 6/(\pi a^2)\sqrt{1-(2r/a)^2}, & \text{if } r \le a/2 \\ 0, & \text{otherwise} \end{cases} && (16\text{–}27) \\ F(\rho) &= 3\sqrt{\pi/2} J_{3/2}(\pi a\rho)(\pi a\rho)^{-3/2} \\ &= \frac{3}{(\pi a\rho)^3}\left[\sin(\pi a\rho) - \pi a\rho\cos(\pi a\rho)\right], && (16\text{–}28) \end{aligned}$$

where a = diameter.

Ring (The brightness is zero except on the circumference.)

$$
\begin{aligned}
f(r) &= \frac{1}{\pi a}\delta(r - a/2), && (16\text{–}29)\\
F(\rho) &= J_0(\pi a \rho), && (16\text{–}30)
\end{aligned}
$$

where a = diameter.

Acknowledgments. This work was supported by the National Science Foundation under grants AST 9117100, AST 9420018, and AST 9802989.

References

Bevington, P. R. 1969, *Data Reduction and Error Analysis for the Physical Sciences*, (New York: McGraw-Hill).

Bevington, P. R. & Robinson, D. K. 1992, *Data Reduction and Error Analysis for the Physical Sciences*, (New York: McGraw-Hill), 2nd edition.

Biretta, J. A., Moore, R. L., & Cohen, M. H. 1986, *ApJ*, 308, 93–109.

Birkinshaw, M. 1998, *Physics Reports*, in press.

Bracewell, R. N. 1986, *The Fourier Transform and its Applications*, (New York: McGraw-Hill), 2nd edition.

Briggs, D. S. 1995, PhD thesis, New Mexico Institute of Mining and Technology.

Bunch, D. S., Gay, D. M., & Welsch, R. E. 1993, *ACM Trans. Mathematical Software*, 19, 109–130.

Carlstrom, J. E., Joy, M., & Grego, L. 1996, *ApJ*, 456, L75–L78.

Conway, J. E., Myers, S. T., Pearson, T. J., Readhead, A. C. S., Unwin, S. C., & Xu, W. 1994, *ApJ*, 425, 568–581.

Flatters, C. 1988, `http://www.aoc.nrao.edu/~cflatter/slime.html`.

Fomalont, E. B. & Wright, M. C. H. 1974, in *Galactic and Extragalactic Radio Astronomy*, ed. G. L. Verschuur and K. I. Kellermann, (Berlin: Springer), 256–290.

Hobson, M. P., Lasenby, A. N., & Jones, M. E. 1995, *MNRAS*, 275, 863–873.

Jennison, R. C. 1958, *MNRAS*, 118, 276–284.

Jones, M., Saunders, R., Alexander, P., Birkinshaw, M., Dillon, N., Grainge, K., Hancock, S., Lasenby, A., Lefebvre, D., Pooley, G., Scott, P., Titterington, D., & Wilson, D. 1993, *Nature*, 365, 320–323.

Kawamura, J., & Masson, C. 1996, *ApJ*, 461, 282–287.

Maisinger, K., Hobson, M. P., & Lasenby, A. N. 1997, *MNRAS*, 290, 313–326.

Masson, C. R. 1986, *Astrophys. J.*, 302, L27–L30.

Myers, S. T., Fassnacht, C. D., Djorgovski, S. G., Blandford, R. D., Matthews, K., Neugebauer, G., Pearson, T. J., Readhead, A. C. S., Smith, J. D., Thompson, D. J., Womble, D. S., Browne, I. W. A., Wilkinson, P. N., Nair, S., Jackson, N., Snellen, I. A. G., Miley, G. K., de Bruyn, A. G., & Schilizzi, R. T. 1995, *ApJ*, 447, L5–L8 .

Parker, R. L. 1977, *Ann. Rev. Earth Planet. Sci.*, 5, 35–64.

Pearson, T. J. 1991, Caltech VLBI analysis programs, *BAAS*, 23, 991–992. (See `http://astro.caltech.edu/~tjp/citvlb/`).

Pearson, T. J. & Readhead, A. C. S. 1984, *ARAA*, 22, 97.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, *Numerical Recipes*, (Cambridge: Cambridge University Press), 2nd edition.

Purcell, G. H. 1973, PhD thesis, California Institute of Technology.

Readhead, A. C. S., Walker, R. C., Pearson, T. J., & Cohen, M. H. 1980, *Nature*, 285, 137–140.

Rogers, A. E. E., Hinteregger, H. F., Whitney, A. R., Counselman, C. C., Shapiro, I. I., et al. 1974, *ApJ*, 193, 293–301.

Schwab, F. R. & Cotton, W. D. 1983, *AJ*, 88, 688–694.

Scott, P. F., Saunders, R., Pooley, G., O'Sullivan, C., Lasenby, A. N., Jones, M., Hobson, M. P., Duffett-Smith, P. J., & Baker, J. 1996, *ApJ*, 461, L1–L4.

Shepherd, M. C. 1997, ed. G. Hunt & H. E. Payne, ASP Conference Series, 125, 77–84. (See `ftp://astro.caltech.edu/pub/difmap`.)

Twiss, R. W., Carter, A. W. L., & Little, A. G. 1960, *Observatory*, 80, 153–159.

White, M., Carlstrom, J. E., Dragovan, M., & Holzapfel, W. L. 1998, *ApJ*, in press (astro-ph/9712195)

Whitney, A. R., Shapiro, I. I., Rogers, A. E. E., Robertson, D. S., Knight, C. A., Clark, T. A., Goldstein, R. M., Marandino, G. E., & Vandenberg, N. R. 1971, *Science*, 173, 225.

Wild, J. P. 1970, *Aust. J. Physics*, 23, 113–115.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

17. Special Problems in Imaging

W. D. Cotton
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A.

Abstract. In practical applications, one or more of the simplifying assumptions that were used in Lectures 1 and 2 to derive the relationships between the interferometer visibility measurements and the image of the sky may be violated. Serious violations of these assumptions result in distortions and/or errors in the image. Practical considerations, such as finite computer resources, may also occasionally create difficulties. This lecture addresses several potential problems from a practical point of view. The general nature of the problems is described, as are the conditions under which they become important. Finally, it discusses techniques that can be used to reduce the distortions and/or the errors introduced into images, and to reduce the computing requirements.

1. Wide Field Problems

This section discusses various common effects that are present to some extent in images of regions of any size, but that become important when a wide field of view is imaged.

1.1. Bandwidth smearing (chromatic aberration)

The effect of finite bandwidth on a correlator was outlined in Lecture 2 and is discussed in detail in Lecture 18, where it is analyzed by expressing u and v as functions of frequency and explicitly averaging over frequency. In general, the effect is to smear $I(l,m)$ locally with a radially oriented image of the bandpass. This smearing is not a proper (position-independent) convolution since the smearing function is a function of (l,m).

When we observe with a finite bandpass $\Delta\nu$, we average the visibilities over a finite region of the (u,v) plane. Smearing occurs when the visibility changes significantly in the region over which the averaging is done, as in Figure 17–1. Since the averaging produced by the bandpass is along a radial line in the (u,v) plane, the smearing in the image plane is also in the radial direction.

A practical example of this effect is shown in Figure 17–2, which shows the image of a bandwidth-smeared extragalactic double source with a weak central component. This observation was made with the VLA at 1.4 GHz with a 50 MHz bandpass, and the source was 12.9′ from the phase tracking center—well outside the plotted field to the lower left.

As described above, the width of the smeared image is proportional to the fractional bandwidth $\Delta\nu/\nu$ multiplied by the separation $\sqrt{l^2+m^2}$ from the delay tracking center. Lecture 18 derives expressions for the beam degradation and reduction in amplitude of a point source for several practical cases. For sufficiently small fields of view, the smearing has less effect than the convolution with the synthesized beam, so it is relatively unimportant. Conservatively, we might consider the image to be substantially distorted if the amplitude on the longest baseline is reduced by more than 5%, or the response is broadened by 5%.

Bandwidth smearing may not be a serious problem if the affected source is not directly of interest but must be imaged only to remove its sidelobes from

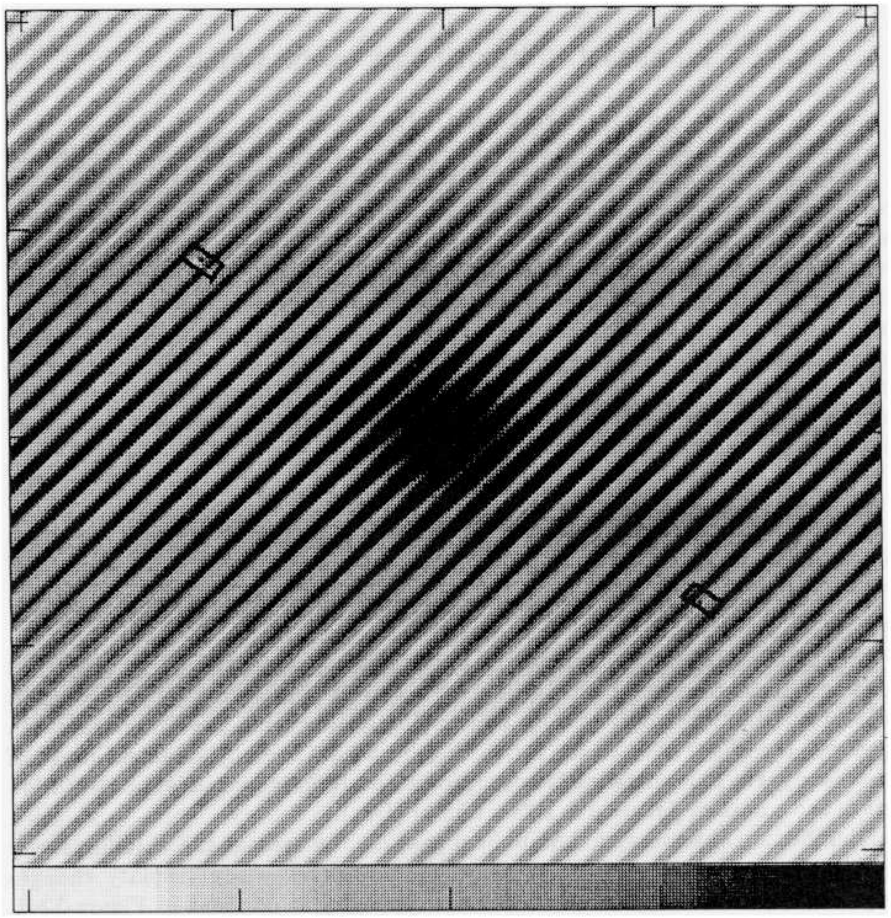


Figure 17–1. The grey scale shows the real part of the inverse Fourier transform (visibility function) of a model source brightness distribution. The boxes indicate the region over which a given data sample might be averaged; the radial extent of the box is determined by the bandwidth, and the azimuthal extent by the time averaging. If the visibility function changes significantly over the region being averaged, as in the case illustrated here, the resulting image will be distorted.

an interesting region closer to the delay tracking center. Bandwidth smearing is a single-valued, symmetric function of u and v, so the observed data correspond to some, rather unlikely, brightness distribution on the sky. The response to the source can therefore be removed by standard deconvolution procedures. Section 1.3 below gives an example of the successful deconvolution of the effects of the source shown in Figure 17–2 from another field.

If an undistorted image is desired, there are several possible approaches to reducing bandwidth smearing. These include: (a) using a single sufficiently

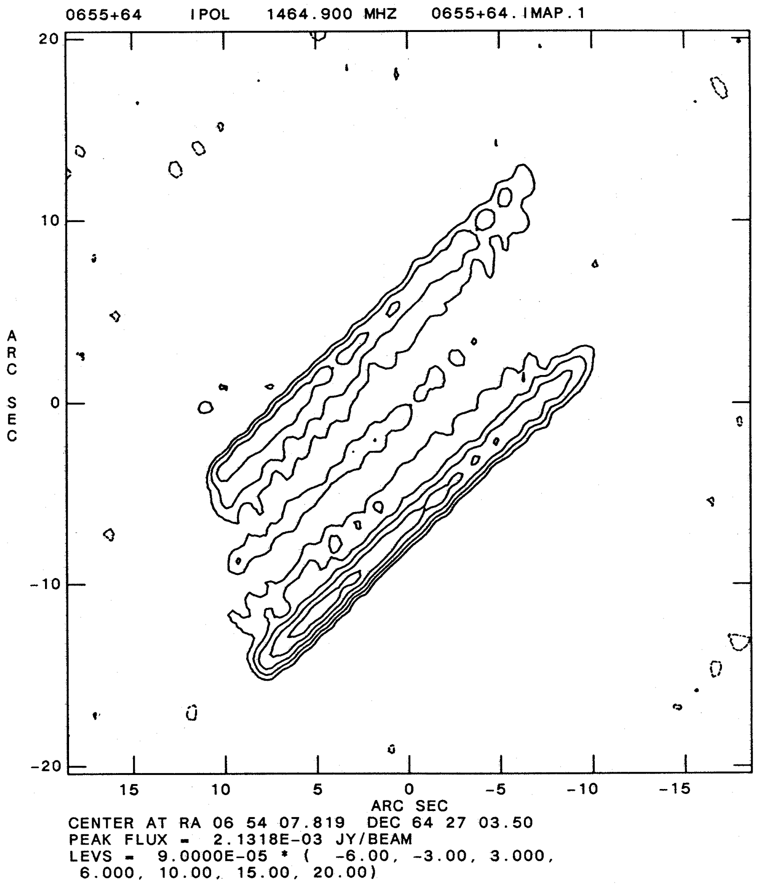


Figure 17–2. The effect of bandwidth smearing on a source 12.9′ northeast of the delay tracking center. The smearing is along the radial direction.

narrow band, (b) narrow bandwidth synthesis,[1] and (c) analytical deconvolution. A related technique, which is not directly used to reduce bandwidth smearing but is sufficiently similar to these methods that it merits attention here, is (d) wide bandwidth synthesis.

1.1.1. Observing with a single narrow bandwidth The effects of bandwidth smearing are proportional to the bandwidth, so the simplest remedy for bandwidth smearing is to observe with a single bandwidth narrow enough that the problem becomes negligible. The resulting sensitivity loss may make this approach unattractive, however.

1.1.2. Narrow "bandwidth synthesis" If the source can be considered to have the same brightness all across the bandpass, then, as in spectral line observing,

[1] The term *bandwidth synthesis*, or *multi-frequency synthesis*, is used to describe the process of improving the (u, v) coverage by independently gridding and combining data obtained in several different frequency channels. See Lecture 21 — *Eds.*

the observing band can be divided up into a number of narrow-band channels—sufficiently many of them that, in each one, bandwidth smearing is no longer a problem. In practice, the requirement for a constant source brightness distribution across the observing band necessitates a small ($\sim$ a few percent) total fractional bandpass.

As was discussed in Lecture 2, Section 10, if each of the narrow-band channels is imaged individually and then averaged, the bandwidth smearing will be that due to the channel bandwidth rather than to the total bandwidth. The individual channels may be combined on a common grid either while gridding (if an FFT is being used) or after making the Fourier transform.

The practical effect of this bandwidth synthesis is that the sidelobes are smeared, rather than the image of the source. This is because explicit use is made of the bandwidth to increase the (u, v) coverage used for the point source response; each of the channels in effect provides its own distinct (u, v) coverage. In many cases, this reduction of the far sidelobe levels will reduce the effects of a distant, strong confusing source better than using the bandwidth smearing to reduce its response.

1.1.3. Analytical deconvolution Several analytical techniques have been suggested for dealing with bandwidth smearing (e.g., Clark 1982b). The principal difficulty with these techniques is that if the image is heavily distorted, then much of the desired information has been lost, and the restoration is likely to tell more about the bandpass functions $g(\nu)$ than about the source.

1.1.4. Wide "bandwidth synthesis" The use of bandwidth synthesis to increase the (u, v) coverage can be expanded to wider bandpasses. The frequency channels need not be contiguous, but may be as widely separated as the electronics will allow; this is a mode frequently used for astrometric and geodetic measurements with very-long-baseline interferometry (VLBI). If the frequency channels are relatively widely spaced (so they span bandwidths of tens of percent), the (u, v) coverage of the observation is significantly improved—which may result in a significant improvement of the quality of the derived image. Unfortunately, in this regime the assumption that the intensity distribution across the source is constant across the bandpass is likely to break down. For these cases the analysis of the data should take into account the variations in the spectral index across the source, and perhaps also spectral curvature. For a more detailed discussion of this technique see Cornwell (1984b).

1.2. Time-average smearing

Time-average smearing is similar to bandwidth smearing, since it is the result of averaging the data over time periods during which the source visibility, on at least some baselines, is not constant. Earth-rotation synthesis arrays use the rotation of the Earth to vary the (u, v) location of the constituent interferometers; thus, the (u, v) locations being sampled are constantly changing. Averaging data over times during which the visibility changes significantly causes an amplitude reduction that will result in a distortion of the derived image.

The effects of time-average smearing are also analyzed in detail in Lecture 18 in terms of the time derivatives of the observing geometry. Due to the complex nature of the effect, its symptoms are not as easily recognized as are those due to

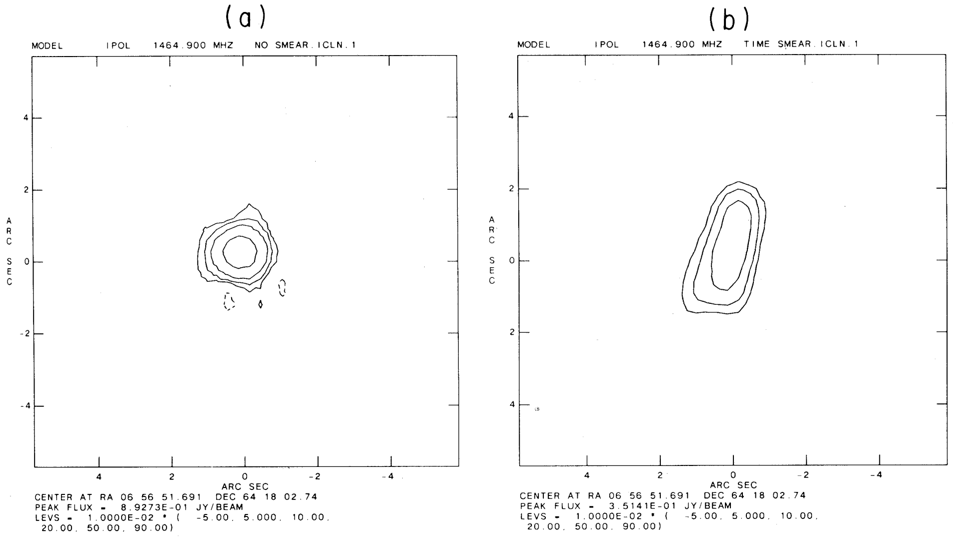


Figure 17–3. **(a)** shows the 'CLEAN'ed image of a point model ~500 synthesized beamwidths west of the phase center without time-averaging, and **(b)** shows the 'CLEAN'ed response to averaged data for the same model, showing the effects of time-average smearing.

bandwidth smearing. However, since longer baselines tend to move more rapidly through the (u, v) plane and to occupy regions of higher spatial frequencies u and v, time-average smearing tends to be stronger on longer baselines. Time-average smearing mimics resolution, and the image of a point source away from the phase center appears resolved and distorted. Since the phase of the response in the (u, v) plane to a source varies increasingly rapidly with increasing separation of the source from the phase center, the extent of the smearing also depends on the separation of the source from the phase center of the pre-averaged data.

If the source is at a celestial pole, then the (u, v) tracks are circular and the smearing is in the azimuthal direction and proportional to the square of the distance of the object from the visibility phase center. In this case, the source image is convolved with the image of the time-averaging function, so the simplest form of data averaging results in the profile of the response becoming rectangular when the smearing is large.

Figure 17–3 shows a severe example of the effects of time-average smearing, using model data. This Figure shows the 'CLEAN' image derived for a given model point source, with and without time-average smearing.

Lecture 18 describes how to check whether time-average smearing is important in a synthesis imaging experiment, and how to compare it to bandwidth smearing. The principal reasons for using long integration times in practise are economic: shorter integration times require more storage medium, more I/O time and more CPU time for the data reduction. If considerations such as these do not dominate, the simplest solution to time-average smearing problems is to use a short integration time, if one is available from the correlator.

If available computer resources dictate some averaging of the data, then there are several approaches. Three of these are (a) baseline-dependent averaging, (b) optimal time series filtering, and (c) multiple fields.

1.2.1. Baseline-dependent averaging As shown above, the effects of time averaging are most severe on the longest baselines. If a given array has a relatively centrally-condensed (u, v) coverage, then much of the data is obtained from the shorter baselines. Thus, the bulk of the data may be significantly reduced in volume if the averaging time is a function of the baseline length, with shorter baselines having longer integration times. In this case, an upper limit to the integration time should be imposed that corresponds to the timescale for instrumental or atmospheric variations, so that self-calibration will be able to remove these effects.

1.2.2. Optimal time series filtering Averaging of data is usually done by convolving a time series of data with a rectangular function and sampling at the center of the function. Recent work in this area suggests that other convolving functions may allow a data compression factor on the order of four using Finite Impulse Response filtering. A good reference is Crochiere and Rabiner (1983). Unfortunately, a convolution on a time sequence (i.e., along a baseline track) does not correspond to a convolution in the (u, v) plane. The effects of other convolving functions, and for that matter the rectangular function currently in use, need further study.

1.2.3. Multiple fields Since the effects of time-average smearing are a function of the separation from the phase center of the pre-averaged data, they can be reduced in a given direction on the sky by shifting the phase center before averaging. Data for multiple fields may be derived from the pre-averaged data by this technique. Unfortunately, multiple copies of the averaged data must be kept. If the data compression due to the averaging is sufficiently large, and the number of fields is sufficiently small, then this technique is practical.

1.3. Sparse fields and confusing sources

Observers are frequently interested in wide fields of view that contain widely scattered sources but which are otherwise mostly empty. This happens either because the sources of interest are widely scattered—e.g., as in surveys—or because there are scattered sources in the field whose sidelobes contribute significantly to the region of interest. (Such sources are usually termed *confusing* sources in radio astronomy). Such fields of view may contain several relatively small, but widely separated regions of interesting emission, with blank sky in between. These regions cannot be deconvolved independently because the sidelobes from one will appear in each of the others.

Figure 17–4 shows an example of the effect of widely scattered confusing sources. This Figure shows the field around the position of a pulsar observed with the VLA at 1.4 GHz. Figure 17–4a clearly shows the sidelobes of distant confusing sources (one of which is shown in Fig. 17–2). To remove the effects of these distant sources by deconvolving the entire region, a 4096 × 4096 image would be necessary.

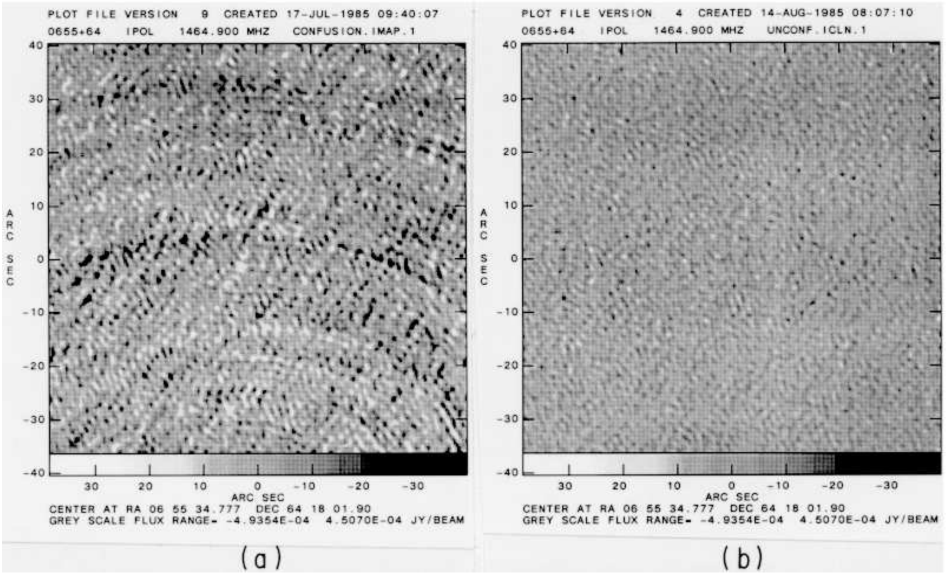

PLOT FILE VERSION 9 CREATED 17-JUL-1985 09:40:07
0655+64 IPOL 1464.900 MHZ CONFUSION.IMAP.1
40
30
20
10
0
-10
-20
-30
-40
ARC SEC
30 20 10 0 -10 -20 -30
ARC SEC
CENTER AT RA 06 55 34.777 DEC 64 18 01.90
GREY SCALE FLUX RANGE= -4.9354E-04 4.5070E-04 JY/BEAM
(a)
PLOT FILE VERSION 4 CREATED 14-AUG-1985 08:07:10
0655+64 IPOL 1464.900 MHZ UNCONF.ICLN.1
40
30
20
10
0
-10
-20
-30
-40
ARC SEC
30 20 10 0 -10 -20 -30
ARC SEC
CENTER AT RA 06 55 34.777 DEC 64 18 01.90
GREY SCALE FLUX RANGE= -4.9354E-04 4.5070E-04 JY/BEAM
(b)

Figure 17–4. **(a)** The region around a pulsar observed with the VLA at 1.4 GHz, showing the sidelobes of distant, confusing sources. **(b)** The same region as in (a) with the effects of the confusing sources removed by 'CLEAN'.

One approach to this problem is to image the entire region and then to restrict the deconvolution to the areas of emission. This approach can be expensive when the image size becomes large, as in the field shown in Figure 17–4. If most of the region to be imaged is blank, then it is more economical to process only the subregions that are of interest.

As the sidelobes of sources in one subregion must be removed from the other subregions, all subregions must be deconvolved in parallel. The 'CLEAN' deconvolution technique is easily adapted to this purpose since it accumulates the deconvolved image by finding and removing a series of delta functions from the image. If the responses to components found in any one subregion are removed from all the others, 'CLEAN' will proceed as though there is a single image with a number of windows.

Figure 17–4b shows the effect on the image from Figure 17–4a of 'CLEAN'-ing four 256 × 256 subregions, centered on the position of interest and three distant confusing sources. The r.m.s. fluctuation in Figure 17–4a is 109 μJy and in Figure 17–4b is 62 μJy. It is of interest to note that the bandwidth-smeared image shown in Figure 17–2 has one of the confusing sources removed; 'CLEAN' properly removed the response, although it could not recover the correct image of the bandwidth-smeared source.

To subtract the sidelobes in the image plane, the dirty beam must be computed for an area twice the size (i.e., four times the area) of the region of interest. Thus, it is often much more economical to subtract the current 'CLEAN' model from the ungridded (u, v) data every so often, then re-grid and re-FFT the data. This approach (termed the Cotton–Schwab algorithm in Lecture 8) is a variant of the Clark modification to 'CLEAN' (Clark 1980) and will be referred to here as the *ungridded subtraction* technique. Other deconvolution methods would similarly benefit by this technique.

Several features of this technique make it attractive for processing single as well as multiple fields of view. The most obvious of these is that the ungridded subtraction allows 'CLEAN'ing (almost) all of an image, rather than only a quarter of its area. Another advantage is that the aliased responses—both to sources outside the subregion and to sidelobes of sources in the subregion which appear outside it—are greatly reduced. Other potential uses of the ungridded subtraction technique will become apparent later.

There are several possible techniques for subtracting a model from the (u, v) data. For 'CLEAN' or other deconvolution techniques that can produce a list of discrete components, a 'direct Fourier transform' can be employed (see Lecture 7). In the more general case, the (inverse) Fourier transform of the model for each field can be computed, and the values at observed (u, v) locations can be interpolated. These methods are discussed below.

1.3.1. 'Direct Fourier transform' The (inverse) so-called 'direct Fourier transform' of a linear combination of N delta functions (point components), evaluated at a given u, v and w, is given by

$$V(u,v,w) = \sum_{i=1}^{N} A_i e^{-2\pi i(l_i u + m_i v + n_i w)} , \qquad (17\text{–}1)$$

where

$$\begin{aligned} A_i &= \text{flux density of component } i\,, \\ (l_i, m_i) &= \text{position of component } i\,, \\ \text{and} \quad n_i &= \sqrt{1 - l_0^2 - m_0^2} \quad \text{with} \quad (l_0, m_0) = \text{center of the field}\,. \end{aligned}$$

The w-term in Equation 17–1 corrects the phase center of the field to the phase center of the (u, v) data, and the sum can be extended over components found in all fields. Similar expressions can be derived for other models (models that include other than point components). The method is relatively efficient when there is a small number of model components or a large number of fields and/or bandwidth synthesis frequency channels, but it may become expensive for large numbers (100,000 or more) of components.

1.3.2. Gridded interpolation Another technique, which becomes attractive when the model cannot be expressed as a manageable number of discrete components, is to compute the (inverse) Fourier transform of the model of a given field and interpolate the model values at the observed (u, v) locations. This process must be done separately for each field, and each frequency channel must be interpolated independently.

1.4. Noncoplanar baseline effects (w-term)

Section 4.2 of Lecture 1 described a small-field approximation to the fundamental Equation 1–5 whereby the transformation became a two dimensional Fourier transform. In the general case this approximation breaks down, and the effects due to ignoring the w-term may become serious.

To estimate the consequences of neglecting the w-term, consider the effect on a point source at (l, m) observed with a single interferometer. As shown in Lecture 2, the phase error (in radians) incurred by ignoring the w-term is:

$$\text{error} \approx \pi w \theta^2 \,, \tag{17–2}$$

where $\theta \equiv \sqrt{l^2 + m^2}$.

If w is a linear function of u and/or v, as in the case of a coplanar array, then the linearly increasing phase error across the (u, v) plane will appear as a position error in the image plane. The apparent position shift is a function of the zenith angle and the azimuth of the source. Thus the source will appear to move during the observations, and the resultant image will show the trace of this apparent motion during the observations. For noncoplanar arrays (e.g., in VLBI) the effect is more complex. This problem has been discussed in a number of other places (Clark 1973c, Hudson 1977, Clark 1981)

For a coplanar array, w in the azimuth of the source is $\sim \sqrt{u^2 + v^2} \sin z$, where z is the instrumental zenith angle. Using this expression for w, Equation 17–2, and the relation "phase error (in turns, i.e., multiples of 2π)" = "position error (radians)" × "spatial frequency (wavelengths)", the apparent position shift in arcseconds is approximately given by:

$$\text{position error} \approx \frac{\theta^2}{2 \times 2.06 \times 10^5} \sin z \,. \tag{17–3}$$

The effects for noncoplanar arrays (e.g., VLBI arrays) will be of the same order of magnitude if the $\sin z$ term is dropped, although, in this case, the effect will not mimic a simple position shift.

If the error derived from Equation 17–3 is small compared to the synthesized beam size, this correction may be ignored. For astrometric or geodetic applications the requirements are more stringent than if only an image is desired. In general, the fields of view imaged with a coplanar array in which w is not zero will be distorted, although the effect can be reduced by restricting the observations as closely as possible to meridian transit.

Examples of the effects of neglecting the w-term in the transform are shown in Figure 17–5. This Figure shows model source data for a point $47\rlap{.}'5$ from the phase center, for VLA (u, v) coverage obtained at 40° declination. Figure 17–5a shows the image derived for a full track of the object, and Figure 17–5b shows the image derived for a single 30 minute subset of the data. Figure 17–5a shows a gross distortion of the image as the apparent position of the source changes during the day. Figure 17–5b appears relatively undistorted, but note the $> 30''$ position error.

There are several techniques for reducing noncoplanarity problems in addition to observing only near the zenith; those which will be discussed here are (a) multiple fields of view, (b) geometric correction, and (c) 3–D FFTs.

1.4.1. Multiple fields of view As was shown above, the errors resulting from ignoring the w-term increase as the square of the distance from the phase center.

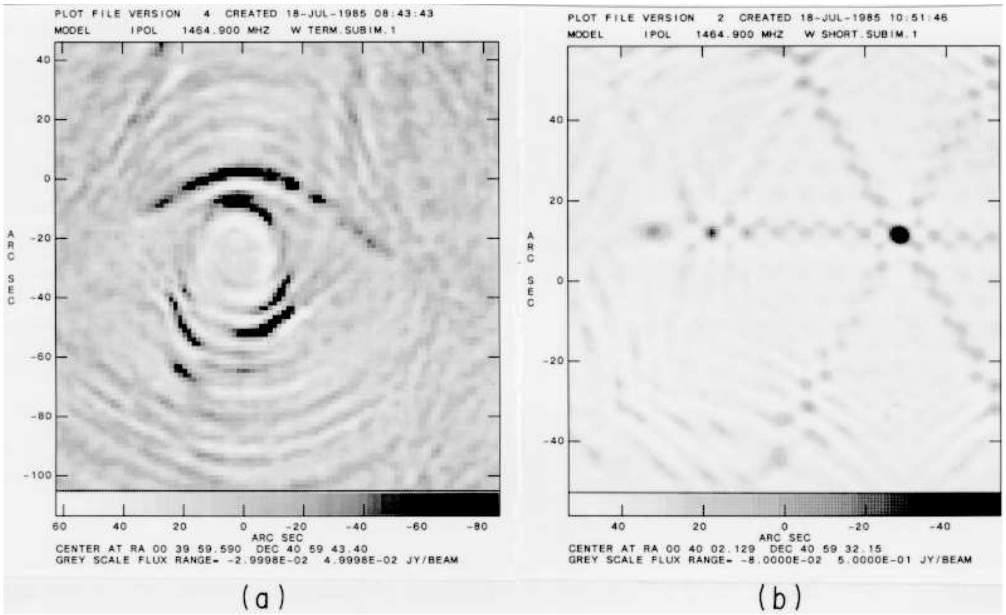


Figure 17–5. **(a)** The response of the VLA to a point model source 47′.5 in RA from the phase center, for full coverage in the VLA **B** configuration at 1.4 GHz. Zeros on the axes label the correct position of the source; the model contained 1 Jy, but the peak in the image is 0.071 Jy. **(b)** Similar to (a), but made using the (u, v) coverage corresponding to only 30 minutes of observation. The peak in the image is 0.948 Jy.

Thus, the errors due to ignoring the w-term can be arbitrarily reduced by breaking the region up into a number of fields of view, each of which is imaged using its center as the phase center. The ungridded subtraction technique discussed previously is useful for deconvolving the resultant images.

1.4.2. Geometric correction If the array is coplanar with nonzero w, or if it can be considered to be so for suitably chosen time intervals, then

$$w = au + bv\,, \tag{17–4}$$

and there will be a simple geometric distortion of the image that can be corrected. This technique, which is especially useful for East–West arrays, is in use at the Westerbork Synthesis Radio Telescope. If the array is only approximately coplanar for intervals of time, then the field can be imaged in each interval, corrected, and (finally) all of the images averaged.

A note is in order here about dividing data into several time segments. Since the Fourier transform is linear, data can be averaged before or after the transform. However, if uniform weighting is being applied to the data, then this correction must be done before the data are divided into time intervals.

1.4.3. 3-D FFTs A more nearly correct, but expensive, method is to do a full three-dimensional FFT and then project the result onto the celestial sphere. This approach is discussed in detail in Lecture 19.

1.5. Nonisoplanatic and antenna polarization effects

A common assumption made during calibration is that the complex gains needed for calibration do not vary with position on the sky. This assumption is unavoidable during the initial calibration phases, since the distribution of signals from the sky is, of course, unknown. This assumption may be incorrect for some wide field observations.

The two principal causes of position-dependent calibration are small-scale variations in the atmosphere, especially the ionosphere, and instrumental—primarily polarization—variations across the antenna pattern. Ionospheric problems become increasingly severe with decreasing frequency, both because the antenna pattern becomes larger and because phase fluctuations become increasingly larger. When the field of view becomes larger than the size of an isoplanatic region (a region over which the phase and amplitude errors induced by the atmosphere can be considered to be constant), position-dependent calibration is required. Position-dependent polarization problems arise in wide field observations when the antenna patterns in the orthogonal polarizations are not identical and/or are not aligned.

By the nature of position-dependent calibration, its application must involve a deconvolution of the image. Schwab (1984b) has suggested a solution to this problem using an adaptation of self-calibration in which the gain at a number of grid points on the sky is determined. The gain at intermediate locations is determined by interpolation. Instrumental gain variation may be computed or accurately measured independently of the observations, but atmospheric effects must be determined from the data.

The corrections can then be applied using an adaptation of the ungridded subtraction technique. The model used to determine the response can incorporate the position and/or time variations in the gain. Several iterations of this technique may be needed.

1.6. Regions larger than the primary beam

It is sometimes necessary to image a region that is large compared with the main lobe of the primary beam pattern $\mathcal{A}(l,m)$ of the array elements. In this case the image must consist of a *mosaic* derived from separate pointings of the array. Since the regions observed by the individual pointings of the array will overlap on the sky, a substantial improvement in the deconvolution may be obtained by deconvolving the regions in parallel. This technique also allows the determination of, and removal of, the effects of relative pointing errors. The analysis must explicitly include the beam pattern $\mathcal{A}(l,m)$ of the array elements; the images of the different regions must also be projected onto the same plane (i.e., have the same tangent point) and must use the same grid of positions on the sky. This technique is discussed in detail in Lecture 20.

2. Time-Variable Effects

There are a number of time-variable effects that are not removed by normal calibration procedures. Two of these, involving variability of the source and of the antenna pattern, are discussed below. In these cases it is frequently desirable to divide the data into short time intervals, but this may have a serious negative impact on the deconvolution of the image. Deconvolution is nonlinear, so combining images after deconvolution is not equivalent to combining them before deconvolution. The dynamic range of the deconvolution depends strongly on the (u, v) coverage used to make the image, so that only a relatively low dynamic range image can be obtained from the short time interval data.

2.1. Variable sources

One of the fundamental assumptions in forming an image using a synthesis array is that the distribution of brightness on the sky remains constant during the observations. If the source varies during the observations, then the image that is derived is not the convolution of the average brightness of the source with the dirty beam derived in the usual manner. This will lead to an incorrect deconvolution for the source. Two classes of violations of the assumption of constancy are considered below.

2.1.1. Variable point sources Pulsars may exhibit considerable brightness fluctuations due to interstellar scintillations, and some compact, galactic sources have been observed to have significant variations on timescales of a day. An example of a deconvolved image derived from data for a time variable point model is shown in Figure 17–6. Various artifacts appearing in this Figure correspond to sidelobes during time periods when the flux density of the source was different from the average. Especially troublesome are the artifacts that appear similar to jets—these are due to the arms of the VLA.

Two approaches that can be taken to the problem of a time-variable point source are (a) to divide the data into time intervals for which the data can be considered to be constant, or (b) to subtract a time-variable point model from the data before making the image. The latter approach is preferable if there is weak extended emission in the field and a high dynamic range image is desired.

2.1.2. Variable extended sources Under some circumstances, extended emission may vary on the time scale of the observations. Two examples of this are observations of the Sun, which can vary on short timescales, and observations of planets, which rotate. In these cases, if an image is desired, then the data must be divided into sufficiently short time intervals. This will result in relatively poor (u, v) coverage and correspondingly poor dynamic range. If the desired result can be described by a time-evolving model, such as for VLBI observations of the rapidly changing galactic object SS 433, then the parameters of the model can be fitted directly to the observations.

2.2. Variable sidelobes

Antennas with altitude–azimuth mounts have the property that the antenna primary beam pattern $\mathcal{A}(l, m)$ rotates on the sky. If there are strong confusing sources outside the main beam of the antenna pattern, they will appear to vary during the observations, as the pattern rotates over them. This is especially problematic at lower frequencies where the primary beam patterns of the array elements are broad and typically contain many strong sources. The effects of

Figure 17–6. The deconvolved ('CLEAN'ed) image derived from model data for a point source with time-variable flux density. The (u, v) distribution used was that of a source observed with the VLA in the **A** configuration at 1.4 GHz.

these sources on the region of interest will not be completely removed by the standard deconvolution techniques.

An approach to this problem is to divide the data into short time intervals and remove the effects of the confusing sources from the data in each interval. After the effects of the confusing sources are removed, the data can be recombined to form the image of the region of interest. For reasons discussed above, the image of the region of interest should not be deconvolved before the different time intervals are combined.

References

Bracewell, R. N. 1978, *The Fourier Transform and Its Applications*, Second Edition, McGraw–Hill, New York.

Clark, B. G. 1973c, VLA Scientific Memorandum No. 107, NRAO.

Clark, B. G. 1980, *AJ*, 89, 377–378.

Clark, B. G. 1981, VLA Scientific Memorandum No. 137, NRAO.

Clark, B. G. 1982b, Lecture No. 10 in *Synthesis Mapping: Proceedings of the NRAO–VLA Workshop held at Socorro, New Mexico, June 21–25, 1982*, A. R. Thompson and L. R. D'Addario, Eds., NRAO (Green Bank, WV).

Cornwell, T. J. 1984b, VLB Array Memorandum No. 324, NRAO.

Crochiere, R. E. & Rabiner, L. R. 1983, *Multirate Digital Signal Processing*, Chapter 4, Prentice–Hall (Englewood Cliffs, NJ).

Hudson, J. A. 1977, Chapter 5, Ph. D. Thesis, The American University, Washington, D.C.

Schwab, F. R. 1984b, *AJ*, 89, 1076–1081.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

18. Bandwidth and Time-Average Smearing

A. H. Bridle & F. R. Schwab
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A.

Abstract.
This is the first of three lectures to deal with problems in imaging wide fields-of-view. Its goal is to quantify the first two effects described in Lecture 17—bandwidth smearing and time-average smearing. Both effects cause the synthesized image to be distorted in ways that cannot adequately be described (except locally) as a convolution of the true sky brightness distribution with a spatially invariant point source response function. Rather, the degree of smearing is a function of angular distance from the delay-tracking center (for bandwidth smearing) or the phase-tracking center (for time-average smearing). The effects therefore persist after simple (position-independent) deconvolution with methods like 'CLEAN' or MEM. Since these distortions cannot be remedied by calibration (or self-calibration), it is important to devise synthesis observing strategies that hold the distortions down to acceptable levels. We wish now to characterize the two effects mathematically and to justify the approximations embodied in the practical formulae used elsewhere.

1. Bandwidth Smearing (Chromatic Aberration)

1.1. General description of the effect

In Lecture 1, the basic Fourier transform relation between the *monochromatic* visibilities V_ν and the *monochromatic* intensity distribution I_ν was given in Equation 1–9:

$$I_\nu(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V_\nu(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{18–1}$$

In practice, although the receiver passbands are of finite width $\Delta\nu > 0$, we treat all the visibility data as though they correspond to measurements at a single central frequency, ν_0. To see how this distorts the synthesized image, consider an infinitesimal bandwidth $d\nu$ centered on frequency ν. The actual spatial frequency coordinates of a visibility for frequency ν are, let us say, (u_ν, v_ν). But when handling the data, we instead assign the frequency-independent coordinates $u_0 = \frac{\nu_0}{\nu}u_\nu$ and $v_0 = \frac{\nu_0}{\nu}v_\nu$, as though all the data had been taken at frequency ν_0. Within any given visibility sample, the data from all incremental bandwidths $d\nu$ within the passband $\Delta\nu$ are averaged, with weights determined by the instrumental passband shape, and are assigned the *same* u_0 and v_0.

How does this *imprecision* in our handling of the (u,v) coordinates affect the computed brightness distribution? We can answer this by using the *similarity theorem* of Bracewell (1978), which states (see also Lecture 7) that if the functions $X(\mathbf{x})$ and $x(\mathbf{u})$ form a Fourier transform pair in n dimensions, i.e., if $X = \mathfrak{F}x$, then rescaling the coordinates in one domain by a factor α corresponds to rescaling the transform in the other domain by the reciprocal scale factor $1/\alpha$, and renormalizing the amplitudes, so that

$$\frac{1}{|\alpha|^n}X\left(\frac{\mathbf{x}}{\alpha}\right) = \mathfrak{F}x(\alpha\mathbf{u})\,. \tag{18–2}$$

In our case, $n = 2$, $\alpha = \nu_0/\nu$, X is to be identified with I, and x with V.

How does the passband *shape* affect the result? In the general case, the i^{th} antenna and its associated electronics would be described by a voltage bandpass characteristic $g_i(\nu')$, where $\nu' = \nu - \nu_0$. The power bandpass of the i–j interferometer pair in an array would then be $G_{i,j}(\nu') = g_i(\nu')g_j^*(\nu')$. We will assume, however, that all antennas and electronic systems are identical, so that the whole array is characterized by a single power bandpass function $G(\nu')$. In this case, which is a major design goal for nearly every synthesis array, the effect of the passband shape can be described by a multiplication in the (u, v) plane, as we now show.

First we rewrite Equation 18–1 in terms of the *bandwidth-smeared* intensity $\widetilde{I}(l, m)$ and the *frequency-dependent* u's and v's:

$$\widetilde{I}(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widetilde{V}(u_0,v_0)e^{2\pi i(u_0 l+v_0 m)}\,du_0\,dv_0\,, \tag{18–3}$$

where the smeared visibilities $\widetilde{V}$ are obtained from the true V's by rescaling, weighting by the passband function $G(\nu')$ and then summing over all infinitesimal bandpasses $d\nu$. In summing the visibilities over frequency, we must take a further important effect into account. The delay-tracking is appropriate to the center of the field-of-view and the center frequency ν_0. For signals at frequency ν arriving from a direction (l, m) at the interferometer with spatial frequency (u_0, v_0), the inserted delay is in error by an amount $\tau = (u_0 l + v_0 m)/\nu_0$, so the phase is shifted by $2\pi(\nu - \nu_0)\tau = 2\pi\nu'(u_0 l + v_0 m)/\nu_0$. The expression for the smeared visibility is therefore

$$\widetilde{V}(u_0,v_0) = \frac{1}{\int_{-\infty}^{\infty} G(\nu')\,d\nu'}\int_{-\infty}^{\infty} V\left(u_0\frac{\nu}{\nu_0}, v_0\frac{\nu}{\nu_0}\right)\left(\frac{\nu}{\nu_0}\right)^2 G(\nu')e^{2\pi i\frac{\nu'}{\nu_0}(u_0 l+v_0 m)}\,d\nu'\,. \tag{18–4}$$

Now consider the bandwidth smearing of a point source with unit amplitude. As Equation 18–4 describes averaging the visibilities V along a *radius* in the (u, v) plane, we can choose a source on the l-axis, at $(l_0, 0)$, with no loss of generality. Using the *shift theorem* for Fourier transforms (Bracewell 1978), the true visibility is

$$V(u,v) = e^{-2\pi i u l_0}\,. \tag{18–5}$$

The array measures this at points described by the sampling function $S(u_0, v_0)$, and the data are multiplied by a weighting function $W(u_0, v_0)$ when making the image (see Sec. 2.2 of Lecture 7). Inserting the sampled and weighted point source visibility into Equation 18–4,

$$\widetilde{V}(u_0,v_0) = \int_{-\infty}^{\infty} S(u_0,v_0)W(u_0,v_0)e^{-2\pi i u_0\frac{\nu}{\nu_0}l_0}\left(\frac{\nu}{\nu_0}\right)^2 G_n(\nu')e^{2\pi i\frac{\nu'}{\nu_0}(u_0 l+v_0 m)}\,d\nu'\,, \tag{18–6}$$

where $G_n(\nu')$ is the *normalized* passband function $G(\nu')/\int_{-\infty}^{\infty} G(\nu')\,d\nu'$. If the fractional bandwidth is sufficiently small, we can put $(\nu/\nu_0)^2 = 1$. Equation 18–3 for the bandwidth-smeared intensity can then be written as

$$\widetilde{I}(l,m) = \tag{18–7}$$
$$\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} S(u_0,v_0)W(u_0,v_0)e^{-2\pi i u_0 \frac{\nu}{\nu_0} l_0} G_n(\nu')e^{2\pi i u_0 \frac{\nu'}{\nu_0} l}\, d\nu'\right] e^{2\pi i u_0 l}\, du_0\, \delta(m).$$

Rearranging the exponentials, we can rewrite Equation 18–8 as

$$\widetilde{I}(l,m) = \tag{18–8}$$
$$\int_{-\infty}^{\infty} S(u_0,v_0)W(u_0,v_0)e^{2\pi i u_0(l-l_0)}\left[\int_{-\infty}^{\infty} G_n(\nu')e^{2\pi i u_0 \frac{\nu'}{\nu_0}(l-l_0)} d\nu'\right] du_0\, \delta(m)\,.$$

This is an interesting form for $\widetilde{I}$—the term in square brackets is the Fourier transform over ν' of the normalized passband function, to an argument that is the delay $\tau = \frac{u_0}{\nu_0}(l-l_0)$ corresponding to position offset $l - l_0$. It is therefore useful to define a *delay function* $d(\tau)$ related to the passband function $G(\nu')$ by

$$d(\tau) = \frac{1}{\int_{-\infty}^{\infty} G(\nu')\, d\nu'} \int_{-\infty}^{\infty} G(\nu')e^{2\pi i \tau \nu'}\, d\nu'\,. \tag{18–9}$$

The effect of finite bandwidth on the measured visibilities can be described as multiplying the true visibilities by this delay function, which depends on both u_0 and $l - l_0$.

Notice also that Equation 18–9 shows that $\widetilde{I}(l)$ is the Fourier transform over u_0 of the product of four functions—the sampling function $S(u_0, v_0)$, the weighting function $W(u_0, v_0)$, the pristine visibility function $e^{-2\pi i u_0 l_0}$, and the delay function $d(\tau)$. Equation 18–9 can therefore be rewritten, using the convolution theorem, as the convolution of four transforms:

$$\widetilde{I}(l,m) = \mathfrak{F}S * \mathfrak{F}W * \int_{-\infty}^{\infty} e^{2\pi i u_0(l-l_0)} du_0\, \delta(m) * \int_{-\infty}^{\infty} d(\tau)e^{2\pi i u_0 l} du_0\,. \tag{18–10}$$

The first two convolutions give us the narrow-band image—a "dirty beam" $B = \mathfrak{F}S * \mathfrak{F}W$ centered on $(l_0, 0)$. The third is a convolution with a *position-dependent* function that we will call the bandwidth *distortion function* $D(l)$. This distortion function is the Fourier transform over u_0 of the delay function $d(\tau)$. Unlike the dirty beam B, the width and amplitude of D vary with radial distance l_0 from the delay center. Furthermore, D is always oriented along the radius to the delay center. The final image $\widetilde{I}$ is therefore a simple, position-independent, convolution $\mathfrak{F}S * \mathfrak{F}W * I * D$ *only* in the trivial case of a single point source I. The bandwidth distortion of an extended image can be thought of as a "radially-dependent convolution".

The above analysis shows that the bandwidth effect can be characterized by three related functions, the (power) *passband function* $G(\nu')$, the *delay function* $d(\tau)$ and the one-dimensional (radial) *distortion function* $D(l)$. We now give explicit forms of these functions, and derive the final "point source response", in a few simple cases. In what follows, we emphasize the radial symmetry by replacing the coordinate l with the radial coordinate $\theta = \sqrt{l^2 + m^2}$ where appropriate.

Just for a moment, let us explicitly include the dependence of the distortion function D on two radial coordinates, θ and θ_0. Then Equation 18–10, expressing the bandwidth smearing effect on the synthesized image $\widetilde{I}$ *locally* as a convolution of four functions, can be recast in a more general form—valid for the entire image—

$$\widetilde{I}(l,m) = \int_0^\infty (B*I)\left(\frac{l\theta_0}{\sqrt{l^2+m^2}}, \frac{m\theta_0}{\sqrt{l^2+m^2}}\right) D(\sqrt{l^2+m^2}-\theta_0, \theta_0)\, d\theta_0 \,.$$

If $D(\theta, \theta_0)$ had no dependence on θ_0, then Equation 18–10 would reduce to a convolution equation, since the distortion function appearing in the above equation would be a function solely of the difference $\theta - \theta_0$.

1.2. Square bandpass, no tapering, square (u, v) coverage

This is the case that was used to illustrate bandwidth smearing in Lecture 17 (see Fig. 17–2).

Passband function:

$$G(\nu') = \begin{cases} 1, & \text{if } |\nu'| < \Delta\nu/2\,, \\ 0, & \text{otherwise}\,, \end{cases} \qquad \text{thus} \quad \int_{-\infty}^{\infty} G(\nu')\, d\nu' = \Delta\nu\,. \qquad (18\text{–}11)$$

Delay function:

$$d(\tau) = \text{sinc}\, \frac{\Delta\nu\,(l_0 u + m_0 v)}{\nu_0}\,. \qquad (18\text{–}12)$$

Distortion function:

$$D(\theta) = \frac{\nu_0}{\Delta\nu\,\theta_0} \Pi\left(\frac{\nu_0\theta}{\Delta\nu\,\theta_0}\right)\,, \qquad \text{where} \quad \Pi(s) \equiv \begin{cases} 1, & \text{if } |s| < \frac{1}{2}\,, \\ 0, & \text{otherwise}\,. \end{cases} \qquad (18\text{–}13)$$

Note that the *width* of this distortion function increases as $\Delta\nu\,\theta_0$, whereas its amplitude decreases as $1/(\Delta\nu\,\theta_0)$. This illustrates the principal characteristics of the bandwidth distortion—reduction in amplitude, and radial broadening, of the point source response. It also illustrates that the two effects preserve the *integrated* flux density of the distorted response.

Sampling function:

$$S(u,v) = \Pi(u/A)\,\Pi(v/A) \qquad (18\text{–}14)$$

(i.e., a filled square of side A, with longest baseline $= A/\sqrt{2}$).

Weighting function:

$$W(u,v) = 1 \quad \text{(uniform weighting)}.$$

Dirty beam:

$$B(l,m) = \text{sinc}\, Al\, \text{sinc}\, Am \qquad (18\text{–}15)$$

The smeared point source response is B_D, the convolution of the dirty beam B (Eq. 18–15) with the distortion function D (Eq. 18–13). At any given offset $\Delta\theta$ from the beam center at θ_0, the amplitude of this response is

$$\begin{aligned} B_D(\Delta\theta, \theta_0) &= \frac{\nu_0}{\Delta\nu\,\theta_0} \int_{\Delta\theta - \frac{\Delta\nu\,\theta_0}{2\nu_0}}^{\Delta\theta + \frac{\Delta\nu\,\theta_0}{2\nu_0}} \text{sinc}\, A\theta\, d\theta \qquad (18\text{–}16) \\ &= \frac{\nu_0}{\pi A \Delta\nu\,\theta_0} \left[\text{Si}\left(\pi A(\Delta\theta + \frac{\Delta\nu\,\theta_0}{2\nu_0})\right) - \text{Si}\left(\pi A(\Delta\theta - \frac{\Delta\nu\,\theta_0}{2\nu_0})\right) \right]\,, \end{aligned}$$

where $\mathrm{Si}(x) \equiv \int_0^x \frac{\sin t}{t}\,dt$ is the sine integral, a standard special function. The HPBW of sinc $A\theta$ is $\theta_{\mathrm{HPBW}} = 1.206/A$. Defining

$$\begin{aligned}
\eta &= \pi A \theta_{\mathrm{HPBW}} = 3.79\,, \\
\alpha &= \frac{\Delta\theta}{\theta_{\mathrm{HPBW}}} = \text{offset from peak response in undistorted HPBW's}\,, \text{ and} \\
\beta &= \frac{\Delta\nu}{\nu_0}\frac{\theta_0}{\theta_{\mathrm{HPBW}}} = \text{fractional bandwidth } \times \text{ radius in HPBW's}\,,
\end{aligned} \tag{18–17}$$

we can rewrite Equation 18–17 to get the following expression for the degraded beam shape $B_D = B * D$ as a function of offset $\Delta\theta$ from the peak response at θ_0 from the delay center,

$$B_D(\Delta\theta, \theta_0) = \frac{1}{\eta\beta}\left(\mathrm{Si}\,\eta(\alpha + \frac{\beta}{2}) - \mathrm{Si}\,\eta(\alpha - \frac{\beta}{2})\right)\,. \tag{18–18}$$

The peak I of the degraded response to the point source is evaluated by setting $\Delta\theta = 0$, so that α is zero, and substituting into Equation 18–18. The fractional reduction in amplitude of the point source due to bandwidth smearing, $R_{\Delta\nu}$ is the ratio I/I_0, where I_0 is the peak response for $\Delta\nu = 0$ (Eq. 18–18 with $\alpha = 0$ and $\beta = 0$). For this case,

$$R_{\Delta\nu} = \frac{I}{I_0} = \frac{2}{\eta\beta}\mathrm{Si}\frac{\eta\beta}{2}\,. \tag{18–19}$$

1.3. Square bandpass, circular Gaussian tapering

For this case, the bandpass, delay and distortion functions are identical to those in the previous example (Eqs. 18–11 through 18–13).

Sampling function:

$$S(u,v) = 1 \quad \text{(over an area large relative to the scale of the taper)}. \tag{18–20}$$

Weighting (tapering) function:

$$W(u,v) = \frac{\sqrt{\pi}\,\theta_{\mathrm{HPBW}}}{\gamma}\exp\frac{-\pi^2\theta^2_{\mathrm{HPBW}}(u^2+v^2)}{\gamma^2}\,, \quad (\gamma \equiv 2\sqrt{\ln 2} = 1.665)\,. \tag{18–21}$$

Dirty beam:

$$B(\theta) = \exp\frac{-\gamma^2\theta^2}{\theta^2_{\mathrm{HPBW}}}\,. \tag{18–22}$$

The calculation of the degraded beam follows that in Section 1.2, leading to

$$B_D(\Delta\theta, \theta_0) = \frac{\sqrt{\pi}}{2\gamma\beta}\left(\mathrm{erf}\gamma(\alpha + \frac{\beta}{2}) - \mathrm{erf}\gamma(\alpha - \frac{\beta}{2})\right)\,, \tag{18–23}$$

where erf is the usual error function. The reduction in amplitude of a point source relative to zero bandwidth is therefore

$$R_{\Delta\nu} = \frac{I}{I_0} = \frac{\sqrt{\pi}}{\gamma\beta}\mathrm{erf}\frac{\gamma\beta}{2}\,. \tag{18–24}$$

1.4. Gaussian bandpass, circular Gaussian tapering

Passband function:

$$G(\nu') = \exp\frac{-\gamma^2\nu'^2}{(\Delta\nu)^2}\,, \quad \text{where } \gamma \equiv 2\sqrt{\ln 2} = 1.665\,, \text{thus} \int_{-\infty}^{\infty} G(\nu')\,d\nu' = \frac{\sqrt{\pi}\,\Delta\nu}{\gamma}\,. \tag{18–25}$$

Delay function:

$$d(\tau) = \exp\frac{-\pi^2(\Delta\nu)^2(l_0u + m_0v)^2}{\gamma^2\nu_0^2}\,. \tag{18–26}$$

Distortion function:

$$D(\theta) = \frac{\gamma\nu_0}{\sqrt{\pi}\,\Delta\nu\,\theta_0}\exp\frac{-\gamma^2\nu_0^2\theta^2}{(\Delta\nu)^2\theta_0^2}\,. \tag{18–27}$$

For this case, the sampling function is again unity, the weighting function is the Gaussian specified by Equation 18–21, and the dirty beam is the Gaussian specified by Equation 18–22. The degraded beam is

$$B_D(\Delta\theta, \theta_0) = \frac{1}{\sqrt{1+\beta^2}}\exp\frac{-\alpha^2\gamma^2}{1+\beta^2}\,, \tag{18–28}$$

and the reduction in the peak response is

$$R_{\Delta\nu} = \frac{I}{I_0} = \frac{1}{\sqrt{1+\beta^2}}\,. \tag{18–29}$$

1.5. Graphs of the main bandwidth smearing effects

Figures 18–1 and 18–2, adapted from Perley (1981a), show how the two main effects of bandwidth smearing vary as functions of the dimensionless parameter $\beta = \frac{\Delta\nu}{\nu_0}\frac{\theta_0}{\theta_{\rm HPBW}}$, for each of the three combinations of band shape and tapering discussed above. The three curves are strikingly similar. When plotted in this way, the variations in undegraded beamwidth due to different (u, v) coverage and tapering are absorbed into the ratio β, emphasizing the utility of this parameter when describing the bandwidth effect.

Note however that the definition of β obscures the true frequency dependence of the bandwidth effect for a *given* synthesis array. Equation 18–9 shows that, for a given array, the delay function depends only on the shape of the passband, not on the observing frequency. The center frequency ν_0 appears in the definition of β only because it multiplies $\theta_{\rm HPBW}$ in the denominator. The product $\nu_0\theta_{\rm HPBW}$ is independent of frequency for a given array. Despite this, it is convenient in practice to factor β in this way.

2. Time-Average Smearing

2.1. General description of the effect

Averaging of the visibility data is another cause of image smearing. Assuming an averaging time τ_a, these averaged data from each correlator are assigned (u, v)

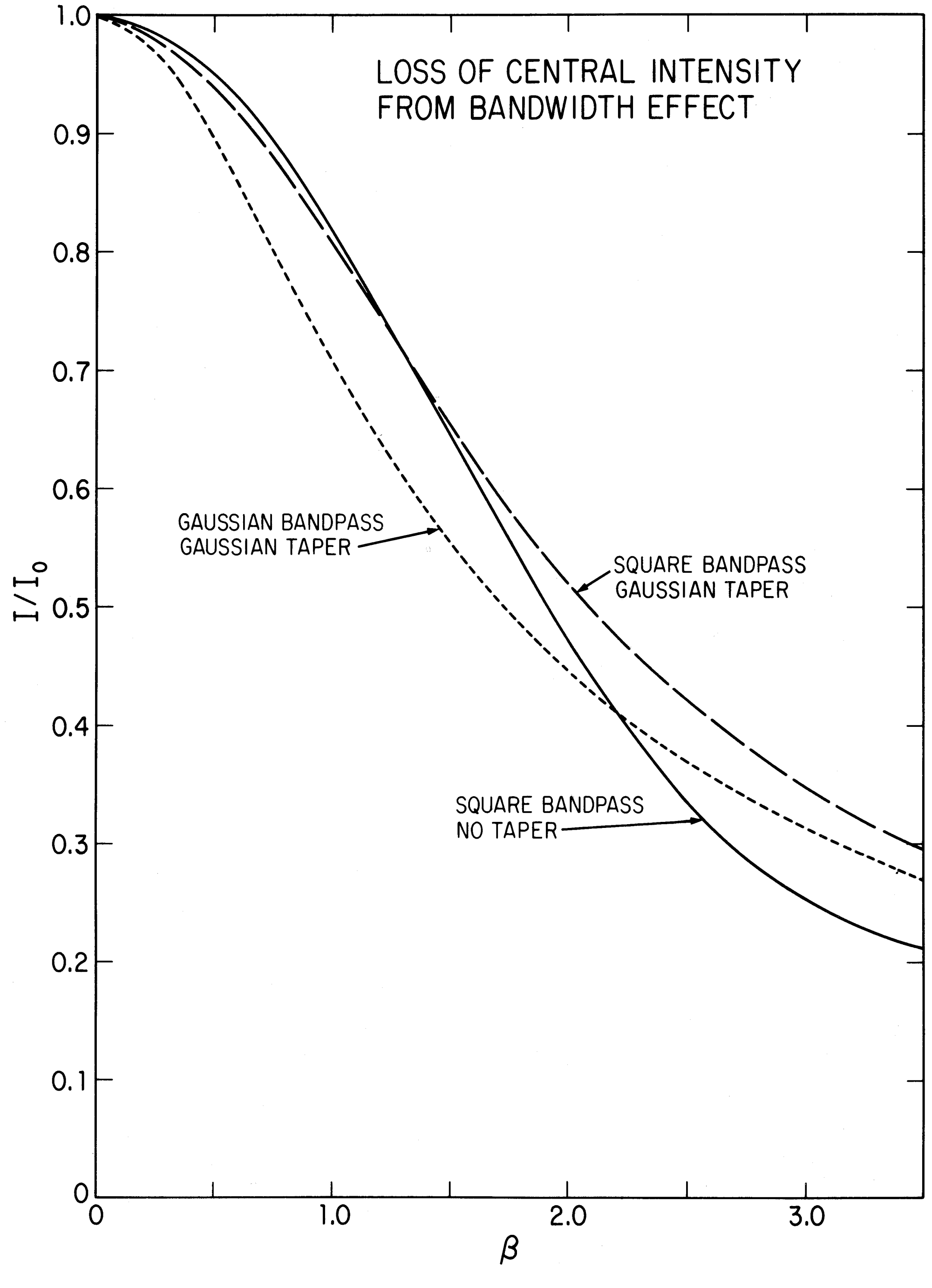


Figure 18–1. The reduction in peak response to a point source, I/I_0, for each of the band shape and taper combinations discussed in Sections 1.2 through 1.4, plotted as a function of the dimensionless parameter β.

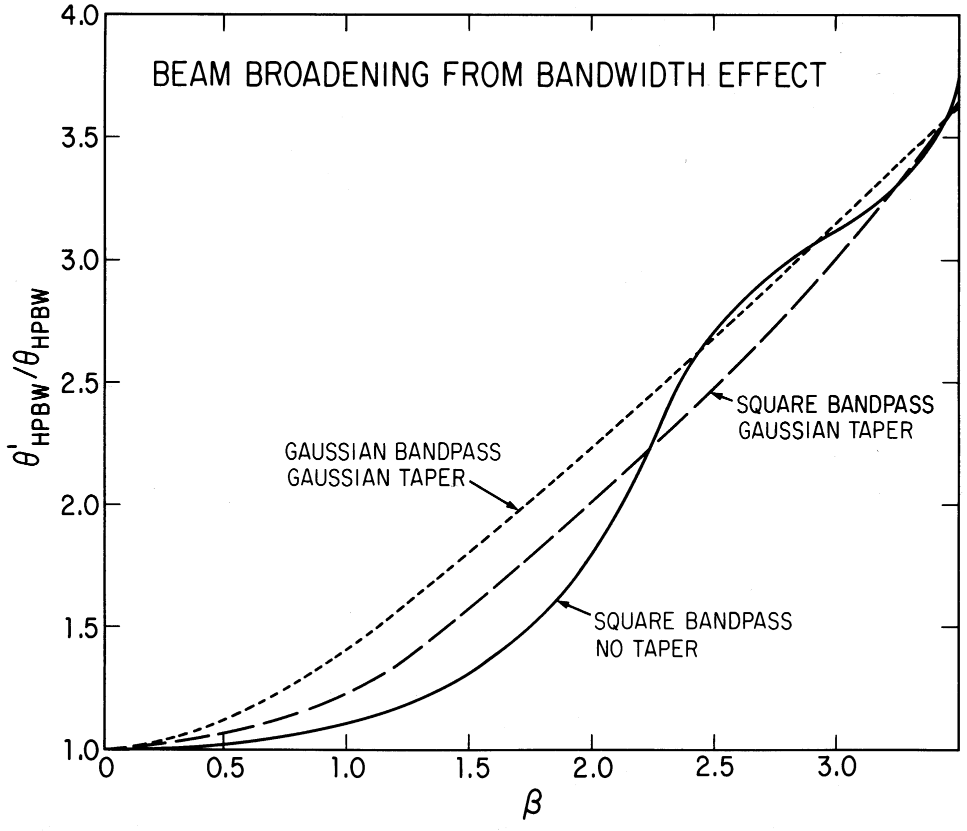


Figure 18–2. The broadening of the point source response relative to zero bandwidth, for each of the band shape and taper combinations discussed in Sections 1.2 through 1.4, plotted as a function of the dimensionless parameter β.

values corresponding to the mid-points t of the averaging intervals, although the data come from time ranges $|\delta t| \leq \tau_a/2$ centered about these mid-points.

For a source at the North or South Celestial Pole, the sampling function is confined to a set of concentric circles in the (u, v) plane, generated by rotating the spacing vector at the Earth's rotational angular velocity ω_e (7.27×10^{-5} rad sec^{-1}). A time offset δt in the assignment of u and v would correspond in this case to a rotation of the visibility function through an angle $\omega_e \delta t$. This would cause the image to be rotated through the same angle, since the Fourier transform commutes with rotations (see Sec. 4.1 of Lecture 7). For an image centered on one of the celestial poles, the effect of time averaging is therefore equivalent to averaging a series of images that are aligned at the l-m origin but have angular offsets up to $\pm\omega_e \tau_a/2$. The weights of the different images in this average reflect the weights of the corresponding times in the time-averaging function. In this particular case, the time-average smearing is therefore equivalent to a distorted *azimuthal* convolution, with a "convolving" function whose shape is determined by the time-averaging function and whose width increases with radius, $\sqrt{l^2 + m^2}$. At the poles, time-average smearing therefore bears an interesting similarity to bandwidth smearing, which produces a distorted *radial* convolution. The general case is, unfortunately, not as simple. It is also more easily understood in terms of the loss of amplitude than in terms of the smearing of the response.

For an object at (l,m) relative to the phase-tracking center, the instantaneous phase is $\phi = 2\pi\nu(ul + vm)$, and the phase rate is therefore

$$\frac{d\phi}{dt} = 2\pi\nu \left(\frac{du}{dt} l + \frac{dv}{dt} m \right) . \tag{18–30}$$

Averaging a waveform of frequency f for a time τ reduces the response by a factor sinc $f\tau$, so for $f\tau \ll 1$ the loss in amplitude is $1 - (\pi f \tau)^2/6$. Integrating the visibility data for a time τ_a therefore reduces the amplitude by a factor

$$R_\tau = \frac{I}{I_0} = \text{sinc} \left(\frac{du}{dt} l + \frac{dv}{dt} m \right) \approx 1 - \frac{\pi^2}{6} \left(\frac{du}{dt} l + \frac{dv}{dt} m \right)^2 , \tag{18–31}$$

valid for small τ_a, where I is the peak response to the source in the image, and I_0 is the peak response in the absence of time-average smearing. We saw in Lecture 2 (Eq. 2–30) that, if L_X, L_Y, and L_Z are the coordinate differences for two antennas, the baseline components (u, v, w) are given by

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} \sin H & \cos H & 0 \\ -\sin\delta\cos H & \sin\delta\sin H & \cos\delta \\ \cos\delta\cos H & -\cos\delta\sin H & \sin\delta \end{pmatrix} \begin{pmatrix} L_X \\ L_Y \\ L_Z \end{pmatrix} , \tag{18–32}$$

where H and δ are the hour-angle and declination of the phase reference position and λ is the wavelength corresponding to the center frequency of the receiving system. We can therefore write

$$\frac{du}{dt} = \frac{1}{\lambda} \left(L_X \cos H - L_Y \sin H \right) \frac{dH}{dt} \tag{18–33}$$

and

$$\frac{dv}{dt} = \frac{1}{\lambda} \left(L_X \sin\delta \sin H + L_Y \sin\delta \cos H \right) \frac{dH}{dt} . \tag{18–34}$$

Substituting these expressions into Equation 18–31 gives an expression for the reduction of amplitude of a point source by time-average smearing on any baseline, as a function of L_X, L_Y, H and δ, provided the reduction is small. The reduction is greatest, for a given baseline, when the apparent diurnal rotation of the sky moves the source at right angles to the fringes associated with that baseline. The reduction is zero when the apparent motion of the source is parallel to the fringes.

2.2. Average effect on an image

For imaging purposes, what is more pertinent is the *average* reduction in amplitude over a 12-hour period. To derive this, we note that $dH/dt = \omega_e$ and use the following relationships for a given baseline (L_X, L_Y, L_Z):

$$\left(\frac{du}{dt}\right)^2 = \frac{\omega_e^2}{\lambda^2}\left(L_X^2 \cos^2 H + L_Y^2 \sin^2 H - L_X L_Y \sin 2H\right) , \tag{18–35}$$

$$\left(\frac{dv}{dt}\right)^2 = \frac{\omega_e^2 \sin^2 \delta}{\lambda^2}\left(L_X^2 \sin^2 H + L_Y^2 \cos^2 H + L_X L_Y \sin 2H\right) , \tag{18–36}$$

and

$$\frac{du}{dt}\frac{dv}{dt} = \frac{\omega_e^2 \sin \delta}{2\lambda^2}\left((L_X^2 - L_Y^2)\sin 2H + 2 L_X L_Y \cos 2H\right) . \tag{18–37}$$

Denoting a 12-hour average by $\langle\ \rangle$, and noting that $\langle \sin^2 H\rangle = \langle \cos^2 H\rangle = \frac{1}{2}$ and $\langle \sin 2H\rangle = \langle \cos 2H\rangle = 0$, we have that

$$\left\langle\left(\frac{du}{dt}\right)^2\right\rangle = \frac{\omega_e^2}{2}\,\frac{L_X^2 + L_Y^2}{\lambda^2} , \qquad \left\langle\left(\frac{dv}{dt}\right)^2\right\rangle = \frac{\omega_e^2 \sin^2 \delta}{2}\,\frac{L_X^2 + L_Y^2}{\lambda^2} ,$$

$$\text{and} \qquad \left\langle\frac{du}{dt}\frac{dv}{dt}\right\rangle = 0 , \tag{18–38}$$

from which we get

$$\langle R_\tau\rangle = \frac{I}{I_0} \approx 1 - \frac{\pi^2}{12}\,\omega_e^2 \tau_a^2 \left(l^2 + m^2 \sin^2 \delta\right) \frac{L_X^2 + L_Y^2}{\lambda^2} . \tag{18–39}$$

Equation 18–39 applies to a single baseline (L_X, L_Y, L_Z). For an array, we can relate the average of the squared lengths of the equatorial projections of the baseline vectors, $\overline{L_X^2 + L_Y^2}$, to the "dirty" half-power beamwidth $\theta_{\rm HPBW}$ by the expression $\overline{L_X^2 + L_Y^2}/\lambda^2 = \alpha/\theta_{\rm HPBW}^2$. The constant of proportionality α is determined by the baseline distribution and by any tapering (weighting) functions applied to the data. For a synthesis image of a source near the North or South Celestial Pole, the average fractional reduction in amplitude $\overline{\langle R_\tau\rangle}$ produced by time averaging for a source a distance θ from the phase-tracking center can therefore be written in the simple form

$$\overline{\langle R_\tau\rangle} \approx 1 - \frac{\alpha\pi^2}{12}\,\omega_e^2 \tau_a^2 \left(\frac{\theta}{\theta_{\rm HPBW}}\right)^2 , \tag{18–40}$$

which is valid in the regime of small intensity losses. We now evaluate the constant α for a few simple cases:

Square coverage, without tapering For square (u, v) coverage of side A (see Eq. 18–14, the beam is given by Equation 18–15, so $\theta_{\rm HPBW}$=1.206/A. For this case, $\overline{L_X^2 + L_Y^2} = A^2\lambda^2/6$, i.e., $\alpha = \frac{1.206^2}{6} = 0.2424$. The average intensity loss factor for a circumpolar point source is therefore

$$\overline{\langle R_\tau\rangle} = 1 - 1.05 \times 10^{-9} \left(\frac{\theta}{\theta_{\rm HPBW}}\right)^2 \tau_a^2 , \tag{18–41}$$

assuming that the loss due to time-average smearing is small.

Circular coverage, without tapering For circular (u, v) coverage of diameter D, the beam has $\theta_{\rm HPBW} = 1.410/D$. For this case, $\overline{L_X^2 + L_Y^2} = D^2\lambda^2/8$, i.e. $\alpha = \frac{1.410^2}{8} = 0.2485$. The average intensity loss factor for a circumpolar point source is therefore

$$\overline{\langle R_\tau \rangle} = 1 - 1.08 \times 10^{-9} \left(\frac{\theta}{\theta_{\rm HPBW}} \right)^2 \tau_a^2 \,, \qquad (18\text{–}42)$$

assuming that the loss due to time-average smearing is small.

Circular coverage with Gaussian tapering If the array produces a Gaussian beam with FWHM $\theta_{\rm HPBW}$ (Eq. 18–22), the (u, v) distribution must approximate its transform (Eq. 18–21), so that $\overline{u^2 + v^2} = \gamma^2/\pi^2\theta^2_{\rm HPBW}$ and $\alpha = \gamma^2/\pi^2 = 4(\ln 2)/\pi^2 = 0.2810$. The average intensity loss factor for a circumpolar point source is therefore

$$\overline{\langle R_\tau \rangle} = 1 - 1.22 \times 10^{-9} \left(\frac{\theta}{\theta_{\rm HPBW}} \right)^2 \tau_a^2 \,, \qquad (18\text{–}43)$$

again assuming that the loss due to time-average smearing is small.

3. Acknowledgments

This lecture draws heavily on previous treatments of these topics by Barry Clark (1981), Rick Perley (1981a) and Dick Thompson (1973; 1982b). We were motivated to add this lecture to the series to collect the results of these earlier treatments, with a few small corrections, into this more widely available format. We hope that in doing so we have amplified them without too much degradation!

References

Bracewell, R. N. 1978, *The Fourier Transform and Its Applications*, Second Edition, McGraw-Hill, New York.

Clark, B. G. 1981, VLA Scientific Memorandum No. 137, NRAO.

Perley, R. A. 1981a, VLA Scientific Memorandum No. 138, NRAO.

Thompson, A. R. 1973, VLA Electronics Memorandum No. 118, NRAO.

Thompson, A. R. 1982b, Lecture No. 5 in *Synthesis Mapping*, A. R. Thompson and L. R. D'Addario, Eds., NRAO (Green Bank, WV).

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

19. Imaging with Non-Coplanar Arrays

Richard A. Perley
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract.
The problem of non-coplanar baselines is discussed along with some techniques for recovering the sky brightness distribution from observations with a non-coplanar array.

1. Imaging in Three Dimensions

1.1. The visibility and image volumes

In Lecture 1, it was shown that the general response of a two-element interferometer to spatially incoherent radiation from the far field can be written as

$$V(u,v,w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(l,m) e^{-2\pi i\left[ul+vm+w\left(\sqrt{1-l^2-m^2}-1\right)\right]} \frac{dl\,dm}{\sqrt{1-l^2-m^2}}\,. \tag{19–1}$$

This expression is appropriate for a phase-tracking interferometer whose bandwidth $\Delta\nu$ is much less than the observing frequency ν. All symbols in Eq. 19–1 are as defined in Lectures 1 and 2, but I have, for convenience, dropped the dependence on the antenna pattern. The general problem is to recover the sky brightness $I(l,m)$ from an ensemble of measurements of the visibility, or coherence, function $V(u,v,w)$. The first two lectures showed that this inversion is simple if either of the following conditions is met:

1. All the measurements of the visibility lie on a plane. In this case, $w = \alpha u + \beta v$, and it is easy to show that Eq. 19–1 reduces to an exact two-dimensional Fourier transform. This condition is met by any one-dimensional interferometer array aligned in the East–West direction; any such array collects data in a plane that is parallel to the Earth's equatorial plane. It is also met by any two-dimensional array if the observation is sufficiently short, as discussed in Section 2.3. (The condition would also be met by a two-dimensional array located at either pole – but this is not a case of great practicality).

2. The field-of-view is limited, by the antenna primary beam for example, to a small enough angular region that the third term in the exponential can be ignored. This approximation also reduces Eq. 19–1 to a two-dimensional Fourier transform, with a small position-dependent error that limits the fidelity of the recovered brightness distribution. I shall return to this in Section 2.1.

Two-dimensional arrays, such as the VLA, cannot often exploit condition 1 (but see Sec. 2.3 for an attempt to do so). Furthermore, at low frequencies, where the primary beam is large, processing the data as if condition 2 holds can lead to severely corrupted images, especially at high resolutions. In the general case, straightforward two-dimensional transformation cannot be used, and a more general inversion technique must be developed.

Before detailing techniques that have been suggested for recovering the sky brightness from the three-dimensional visibility function, it is useful to define the *image volume* and to show its relationship to the visibility function. It turns out that there is a pretty geometrical interpretation of this relation that can help in visualizing the general problem.

To simplify the following development, first recast Eq. 19–1 by removing the phase-tracking, i.e., multiply through by $e^{-2\pi i w}$, to give

$$V'(u,v,w) = \iint I(l,m) e^{-2\pi i \left[ul+vm+w\sqrt{1-l^2-m^2}\right]} \frac{dl\, dm}{\sqrt{1-l^2-m^2}} \,. \tag{19–2}$$

In this equation, and in all subsequent integrals that appear in this lecture, the limits to the integration are from minus infinity to plus infinity. Note first that this is *not* a Fourier transform relation between the visibility, $V'(u,v,w)$, and the sky brightness, if only because the visibility is a function of three variables (the baseline coordinates u, v and w), while the sky brightness is a function of two (the direction cosines l and m). Equation 19–2 closely resembles a Fourier relation, however. Indeed, introducing the direction cosine $n = \sqrt{1-l^2-m^2}$ makes Eq. 19–2 seem tantalizingly close to a three-dimensional relation.

Emboldened by this, I shall take the three-dimensional Fourier transform of $V'(u,v,w)$, and then relate the result to the two-dimensional sky. To do this, I must relate the function

$$F(l,m,n) = \iiint V'(u,v,w) e^{2\pi i(ul+vm+wn)}\, du\, dv\, dw \,, \tag{19–3}$$

where V' is defined as in Eq. 19–2, to the sky brightness $I(l,m)$. Note that in this expression, the quantity n is considered an *independent* variable, even though it is really a function of the direction cosines l and m. The implications of this will be clearer later.

Obtaining the desired relation is straightforward, although tedious. Substituting Eq. 19–2 into Eq. 19–3 gives

$$\begin{aligned} F(l,m,n) \quad = \quad & \iiint \left\{ \iint \frac{I(l',m')}{\sqrt{1-l'^2-m'^2}} e^{-2\pi i(ul'+vm'+w\sqrt{1-l'^2-m'^2})}\, dl'\, dm' \right\} \\ & e^{2\pi i(ul+vm+wn)}\, du\, dv\, dw \,, \end{aligned}$$

where I have introduced the dummy variables l' and m' to simplify the notation. This expression can be manipulated to give

$$\begin{aligned} F(l,m,n) = \iint \Bigg\{ \iiint & \frac{I(l',m')}{\sqrt{1-l'^2-m'^2}} \\ & e^{-2\pi i u(l'-l)} e^{-2\pi i v(m'-m)} e^{-2\pi i w(\sqrt{1-l'^2-m'^2}-n)}\, du\, dv\, dw \Bigg\}\, dl'\, dm' \,. \end{aligned}$$

The integrals over u, v, and w can be evaluated using the general result

$$\delta(l'-l) = \int e^{-2\pi i u(l'-l)}\, du \,, \tag{19–4}$$

to give the expression:

$$F(l,m,n) = \iint \frac{I(l',m')}{\sqrt{1-l'^2-m'^2}} \delta(l'-l)\delta(m'-m)\delta(\sqrt{1-l'^2-m'^2}-n)\, dl'\, dm'. \tag{19–5}$$

These integrals are elementary, giving

$$F(l,m,n) = \frac{I(l,m)\delta(\sqrt{1-l^2-m^2}-n)}{\sqrt{1-l^2-m^2}}\,. \tag{19–6}$$

Equations 19–3 and 19–6 relate the direct three-dimensional Fourier transform of the analytic visibility $V(u,v,w)$ to the two-dimensional sky brightness $I(l,m)$. They were first found by Barry Clark. They justify calling the function $F(l,m,n)$ the three-dimensional *image volume*, so I denote it hereafter by $I(l,m,n)$.

1.2. Interpretation of these formulae

It is useful at this point to contemplate the meaning of Eq. 19–3. It says that there is a three-dimensional Fourier transform relation between the three-dimensional coherence function $V(u,v,w)$ and a three-dimensional image volume $I(l,m,n)$. This volume is not to be associated with physical space (in which the third coordinate would be depth), because its coordinates l, m, and n are *direction cosines.* Equation 19–6 shows that coordinates with radius other than one are in a non-physical part of the image volume—the only non-zero values of I lie on the surface of the sphere of unit radius. The sky brightness distribution is a function of two variables, and the third variable n is introduced solely to establish a formal Fourier relation. The image volume $I(l,m,n)$ is a function of three variables, but the only physically meaningful quantities within it lie on the sphere of unit radius defined by $n = \sqrt{1-l^2-m^2}$.

Equations 19–3 and 19–6 allow us to use the wonderful range of properties of the Fourier transform to explore the general wide-field imaging problem. In particular, note that these relations implicitly used continuously defined functions, rather than discretely-sampled ones—I assumed full knowledge of the coherence function $V(u,v,w)$ for all spacings. In practice, we know only the *sampled* coherence function, $S(u,v,w)V(u,v,w)$, where $S(u,v,w)$ is the sampling function, describing the location of the samples of the visibility (Lecture 7). The convolution theorem can then be used to show that the result of the Fourier inversion becomes

$$I_d = I * B_d\,, \tag{19–7}$$

where the $*$ operator denotes three-dimensional convolution, and $B_d(l,m,n)$ is the transform of the sampling function (i.e., a *three-dimensional "dirty beam"*). Thus the result of Fourier transforming the non-phase-tracked visibilities is the sky brightness convolved by a beam, just as in two dimensions. But the locations of the brightness maxima that correspond to real structure on the sky are constrained to lie on the unit sphere, rather than on a plane.

Figure 19–1 shows a geometric interpretation of these results. It displays a cut through the image volume along the $(l, m=0, n)$-plane. The unit sphere is shown as a unit circle, and sources of radiation at various values of l are

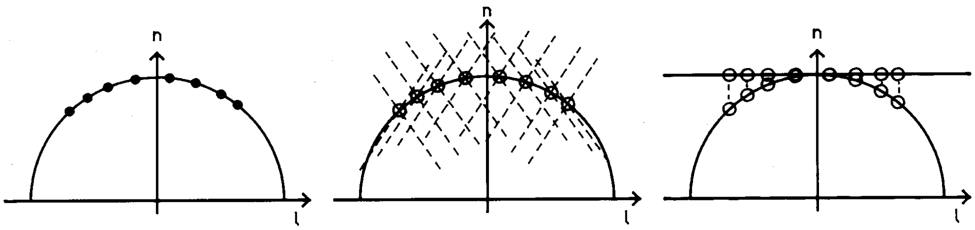


Figure 19–1. The image volume and its relation to the sky brightness. *(Left)* Three-dimensional transformation of the analytic visibility function maps the sky brightness onto a unit sphere. The dots represent these sources. *(Middle)* Convolution with a dirty beam results in sidelobes, shown as dashed lines, throughout the volume above and below the unit sphere. *(Right)* After deconvolution, the images are represented by finite-size "clean beams" on the unit sphere. The two-dimensional image is recovered by projection onto the tangent plane, indicated by vertical dashed lines.

shown as dots along this circle. The n-axis points to the center of the field of interest. (Recall that the phase tracking is turned off, so I cannot call this the "phase-tracking center". It retains its geometric significance, though, since the baseline components are calculated with respect to this direction.) Because the measurements of the visibility are finite in number and extent, the computed image is the convolution of this representation of the sky with the dirty beam. This is shown in Figure 19–1*b*. Note that the volume inside and outside the unit circle is no longer empty, but contains the summed sidelobes of the physical features along the unit circle. Because Eq. 19–7 is a true convolution (if the bandwidth is sufficiently narrow and the effects of time-averaging are small—see the contribution by Alan Bridle in these proceedings), the three-dimensional dirty beam can be deconvolved from the image volume by standard techniques such as 'CLEAN'. Figure 19–1*c* depicts the result—the "dots" of Figure 19–1*a* are replaced by three-dimensional 'CLEAN' beams. Of course, our final goal is a two-dimensional image. The derivation of this is illustrated in Figure 19–1*c*, where the output image is shown as the projection onto the (l, m) plane of the 'CLEAN'ed three-dimensional image.

A better understanding of the relationship between the images formed in this '3-d' space, and the projection onto the tangent '2-d' image plane can be obtained by considering 'snapshot' observations, as illustrated in Figure 19–2. A snapshot observation with a two-dimensional array such as the VLA produces a data set which is coplanar, so the resulting image in '3-d' space consists of the convolution of the true distribution on the unit sphere with a 'ray beam' which thus projects the distribution onto the tangent plane at an angle determined by the array geometry at the time of observation, as illustrated in Figure 19–2*a*. At a different time, the array geometry has changed, so the projection onto the tangent plane occurs at a different angle, as illustrated in Figure 19–2*b*. Since any integrated observation can be considered a summation of snapshots, it can be understood that the resulting image on the tangent plane due to a long integration is summed projection of the true brightness distribution on the unit sphere onto the tangent plane with a set of rotating 'ray beams', whose point of rotation is fixed to the locations of each object on the unit sphere, as illustrated in Figure 19–2*c*. Thus, the image of any real object on the tangent plane will

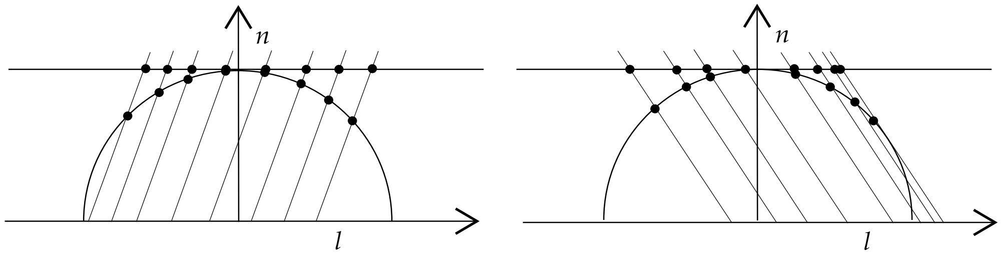


Figure 19–2. The image volume and its relation to a 'standard' two-dimensional image. *(Left)* At a particular time, a 'snapshot' with a two-dimensional array will project the true structure on the unit sphere onto the tangent plane with a 'ray beam', tilted at a particular angle given by the geometry of the array at the time of observation. *(Right)* At a later time, the array geometry has changed due to earth rotation, so the projection is now at a different angle. The apparent positions of the objects which are not located at the tangent point have changed with respect to the earlier observation.

rotate in time, with the sum effect being a distorted image whose centroid is generally in the wrong position.

Analysis of the projected loci is given in Section 2.3 of this lecture.

It may be objected that this formalism produces an unphysical result—namely that the brightness, a real physical quantity, is represented by an unphysical delta function. It is, however, no less physical than the two-dimensional analog given in Lectures 1 and 2, where the sky brightness is represented on a plane of zero thickness. The difference between the plane and the sphere is one of geometry only. Further, recall that the integral of the delta function is well-defined, so that the projection of this representation onto the l,m plane gives a physically meaningful result. Finally, note that the presence of the delta functions is a direct result of the use of the coherence function, which is assumed to be continuously defined and known for all separations. As I have discussed above, the practical application of these results involves finitely many samples of the coherence, so that the derived images are convolved by the dirty beam. Thus, the images have a "thickness" that can be sampled and manipulated, as with two-dimensional images.

In practice, we would prefer not to use the phase-unrotated data. The above derivation used this version of the visibilities only to simplify the mathematics. The effect of using the data produced from a phase-tracking interferometer is a straightforward application of the shift theorem of Fourier transforms. The phase-tracked visibilities are related to the raw visibilities by a phase rotation: $V(u,v,w) = e^{2\pi iw}V'(u,v,w)$. The image volume is therefore shifted by one unit in the conjugate variable, i.e., along the n-axis; the image sphere is shifted so that the phase tracking center lies at the origin of the image volume. This process is illustrated in Figure 19–3a and 19–3b, showing the two-dimensional image circle before and after phase-shifting by the w-phase.

1.3. Direction cosines and image coordinates

The preceding discussions are all couched in terms of the direction cosines. It is obviously desirable to express the output images as functions of the astronomical

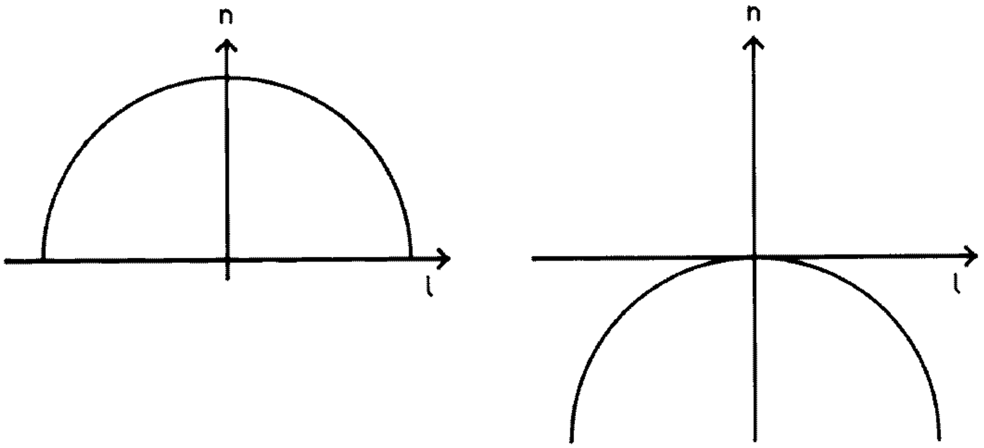


Figure 19–3. An illustration of the effect of phase-tracking on the image volume. The left side shows the image formed by three-dimensional Fourier inversion of the data without phase-tracking. The image circle is centered on the (l, m, n)-coordinate origin. The right side shows the image circle after Fourier inversion of phase-tracked data. The phase-tracking center is now located at the coordinate origin.

coordinates (α, δ). For reference, I now give the relations between these,

$$l = \cos\delta \sin\Delta\alpha\,, \tag{19–8}$$

$$m = \sin\delta\cos\delta_0 - \cos\delta\sin\delta_0\cos\Delta\alpha\,, \tag{19–9}$$

$$n = \sin\delta\sin\delta_0 + \cos\delta\cos\delta_0\cos\Delta\alpha\,. \tag{19–10}$$

In these expressions, $\Delta\alpha = \alpha - \alpha_0$, and (α_0, δ_0) are the coordinates of the phase-tracking center. These equations can be inverted to allow conversion from (l, m) to (α, δ):

$$\delta = \arcsin\left(m\cos\delta_0 + \sqrt{1 - l^2 - m^2}\,\sin\delta_0\right)\,, \tag{19–11}$$

$$\Delta\alpha = \arctan\left(\frac{l}{\sqrt{1 - l^2 - m^2}\,\cos\delta_0 - m\sin\delta_0}\right)\,. \tag{19–12}$$

The values of l and m are derived from the image by multiplying the cell offset (a number) by the cell size (an angle).

1.4. Some fine points

The w-term is often considered to be a delay, so that one might conclude that a spectral-line correlator (which produces a lag spectrum by varying the delays before correlation) might be used to sample the w-coordinate better. This is wrong. The origin of the third term in the phase factor of Eq. 19–1 is geometric, and has nothing to do with delay. To understand the origin of this term more clearly, refer to Figure 19–4.

Figure 19–4 shows two elementary interferometers observing the same object. For simplicity, I have arranged the phase-tracking center to be vertical, and shall consider only two baseline components, u and w. The only difference

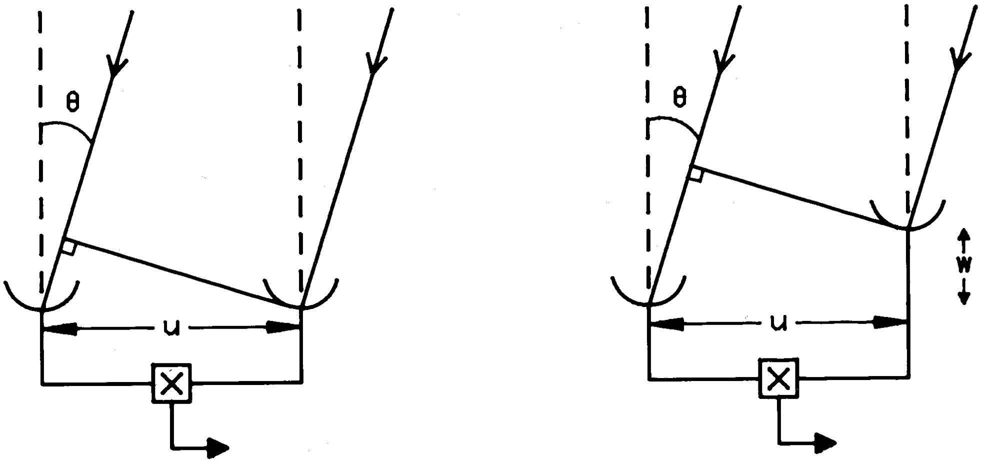


Figure 19–4. An example showing the geometric origin of the third phase term. The left side shows a level interferometer, observing a source of radiation from angle θ. The right side shows an interferometer of the same projected spacing u (with respect to the phase-tracking center), but with the addition of a vertical baseline component, w. The phase of the same source, with respect to the phase-tracking point, is different.

between the two interferometers is their geometry—the first is built on a level plain, the second is (say) on the side of a mountain, so that the baseline has a large w-component. We are interested in the phase of a signal arriving from angle θ with respect to that of a signal arriving from the phase-tracking center.

Consider first the level interferometer. Here the phase of the signal from the reference position is 0, while that from the object at angle θ is

$$\phi_{\text{level}} = 2\pi ul\,, \tag{19–13}$$

where u is the baseline in wavelengths and $l = \sin\theta$ is the direction cosine. The same expression gives the relative phase.

Now consider the tilted interferometer. The reference phase is $\phi_{\text{ref}} = 2\pi w$ and the phase of the signal arriving from angle θ is $\phi = 2\pi(w + u\tan\theta)\cos\theta$. Thus, the relative phase is $2\pi[w(\cos\theta - 1) + u\sin\theta]$. In terms of direction cosines, we have

$$\phi_{\text{tilt}} = 2\pi\left[ul + w\left(\sqrt{1-l^2} - 1\right)\right]\,. \tag{19–14}$$

The apparent phase of an off-axis source is changed by the changed geometry of the interferometer. Because of this, there can be cases in which using only the projected components of the baseline for imaging (i.e., two-dimensional transformations) causes errors. The manifestations of this dependence of the source phase on the baseline geometry are distortions in the 2–D image.

Nevertheless, a spectral line correlator can still enhance wide-field imaging, because transforming the lag spectrum recovers the frequency spectrum of the visibility. For continuum emission, this is equivalent to gaining new visibilities at different effective baselines, so that the visibility volume (which is sparsely filled under the best of conditions) is better sampled.

Use of finite bandwidths broadens the response to point sources in two-dimensional imaging. What is the effect in three? In Lecture 18 it was shown

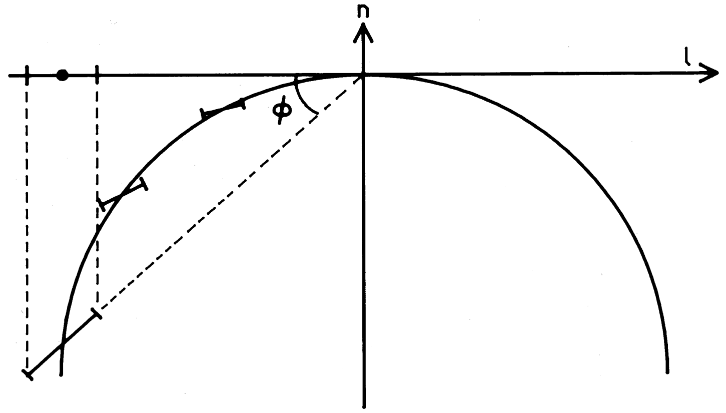


Figure 19–5. An illustration of the bandwidth broadening effect in three-dimensional imaging. The broadening is proportional to the distance from the coordinate origin, and aligned along the radius from the origin. The projection of this broadening onto the tangent plane gives the observed response.

that the effect of using a wide bandwidth is conveniently thought of as multiplication by a delay function in the (u, v) plane, so that emission away from the delay-tracking center contributes a phase slope across the (u, v) plane. Because a continuum correlator integrates the visibilities over a region whose linear size is proportional to the fractional bandwidth along a radius in the (u, v) plane, there is a significant loss of response if the visibility changes within this averaging window. The effect of the bandwidth locally is to convolve the true source structure with a distortion function that has the shape of the bandpass and whose width is given by $\sqrt{l^2 + m^2}\,(\Delta\nu/\nu)$. That is, the broadening is proportional both to fractional bandwidth and to the radial distance from the delay-tracking center.

The same approach can be applied to the visibility cube. I have noted that the image corresponding to phase-tracked data is defined on the surface of a sphere, with the phase-tracking center located at the point where the image plane touches the top of the sphere. Sources located on the sphere will be represented in the visibility volume by phase gradients, just as in two dimensions. Following the analysis in Lecture 18, it is found that the broadening is radial, and is proportional to the distance from the origin and to the fractional bandwidth. The geometry is given in Figure 19–5.

The distance from the origin to an object with coordinates (l, m) is $r = 2(1 - \sqrt{1 - l^2 - m^2})$. The angle of the broadening is given by $\cos\phi = \sqrt{l^2 + m^2}/r$ so that the broadening of the projected image is proportional to $r\cos\phi\Delta\nu/\nu$, or $\sqrt{l^2 + m^2}\,(\Delta\nu/\nu)$. Thus, the effective broadening on a projected image is the same as on a two-dimensional image.

2. Techniques for Recovering the Source Brightness

This section discusses some techniques that have been proposed for recovering the sky brightness from visibility data obtained with a non-coplanar array. Most of them have not been tested—they are "paper solutions".

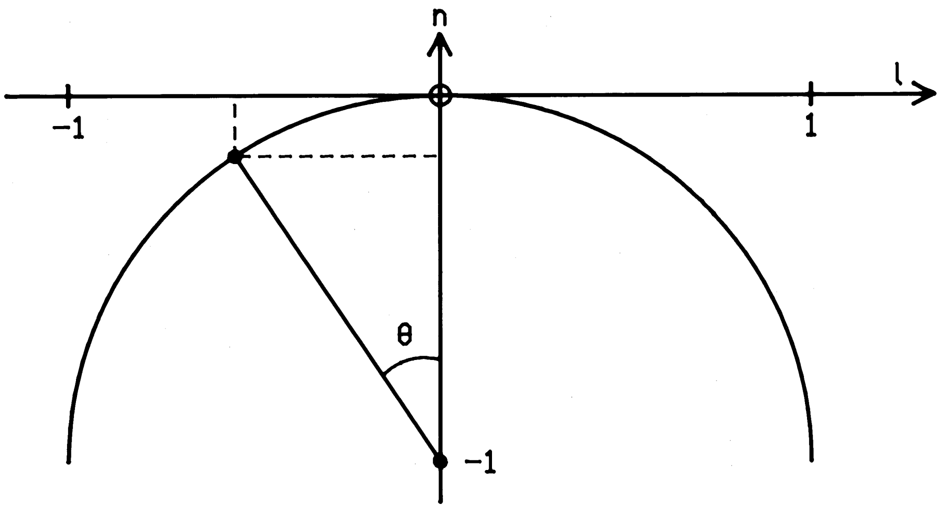


Figure 19–6. A schematic illustration of the distance between the tangent plane and a source located on the image sphere. The number of horizontal planes required is about twice the number of synthesized beams between the tangent plane and image sphere.

2.1. Three-dimensional Fourier transform

Since the three-dimensional transform of the visibility function implies that a three-dimensional volume contains the sky brightness on the surface of a sphere of unit radius, the most straightforward method of recovering the sky brightness is to transform the data directly in three dimensions. The simplest approach would be to grid the data into a visibility cube, then perform a three-dimensional FFT. It turns out, however, that in most applications the number of planes required on the n-axis is small, so that using an FFT in this dimension will produce severe aliasing. An ungridded transform (the so-called 'direct Transform') over this dimension seems prudent.

An obvious question when making a three-dimensional image is: "how many n-planes are required?" The answer comes from considering Figure 19–6. This shows a plane through the image cube, the inscribed circle representing the unit sphere. The plane tangent to the sphere is the image resulting from a standard two-dimensional transformation of the (phase-tracked) visibility data. The image appearing on each plane is the result of convolving the sky brightness with the three-dimensional dirty beam. The size of this beam (assuming a full synthesis in which the data are taken from horizon to horizon) in the n-direction is about the same as in the l- and m-directions, namely: $\theta \approx \lambda/B_{\max}$ where $B_{\max}$ is the longest baseline. Clearly, the n-dimension of the beam must be sampled at least as often as the l- and m-dimensions, that is, at separations no greater than $\delta n \approx \lambda/2B_{\max}$. Figure 19–6 shows the n-component of the distance between the tangent plane and a point separated by an angle θ from the phase-tracking center. This component is simply $n_d = 1 - \cos\theta \approx \theta^2/2$. For critical sampling, the number of planes required on the n-axis is

$$N_{\text{planes}} = n_d/\delta n = B\theta^2/\lambda\,. \tag{19–15}$$

At frequencies below about 1 GHz, this angle will usually be the full primary beamwidth, since the dynamic range is limited by the sidelobes of the background sources. In this case, $\theta = \lambda/D$, so

$$N_{\text{planes}} = \lambda B/D^2 \, . \tag{19–16}$$

This equation, in words, reads "The number of planes required on the n-axis equals the field-of-view in radians times the field-of-view in synthesized beamwidths."

Another useful rule-of-thumb is the following: if the two-dimensional image required to describe the object, or field-of-view, contains N_l cells, and if the full angular size of the object is θ_F, then the number of required horizontal planes is

$$N_{\text{planes}} = N_l \theta_F / 8 \, . \tag{19–17}$$

Table 19–1 shows the number of planes required in the n-direction to allow imaging of the full primary beam (roughly, to the 10 dB points) for all VLA configurations, as a function of observing wavelength.

Table 19–1
Number of n-axis Planes Required for 3–D Imaging

	— VLA Configuration —			
Wavelength	**A**	**B**	**C**	**D**
4 m	225	68	23	7
1 m	56	17	6	2
20 cm	11	4	2	1
6 cm	4	2	1	1

The analysis in this preceding section is a worst case, corresponding to observing the object from horizon to horizon. A more realistic case will take into consideration that the observations will begin and end at an elevation significantly greater than zero. If, for example, the observations are taken such that the maximum delay (expressed in meters) is a factor f less than the baseline (*i.e.* $w_{\max} = B_{\max}/f$), then the number of required vertical planes is reduced by the same factor, since the depth of the 3-d beam has been lengthened by this factor. It should be clear that the factor f is given by $f = 1/\cos(E_{min})$.

Above about 5 GHz, the field-of-view to be processed will usually be set by the size of the target object, since the primary beam will effectively remove background objects. Those sources whose sidelobes are important can be efficiently removed through use of (u, v) subtraction.[1] Thus, the full three-dimensional deconvolution does not often have to be made. However, this is not the case at lower frequencies. Here, the primary beams of typical array elements are large, and the number of background sources becomes so great that the noise on any image is dominated by the sidelobes of the undeconvolved background objects.[2]

[1] As performed, for example, by the AIPS programs 'IMAGR' or 'UVSUB'.

[2] For the VLA at 327 MHz, this noise limit is between 1 and 10 mJy, far higher than the thermal noise limit for almost all observing. Thus, at such low frequencies, full three-dimensional deconvolution is necessary for all VLA configurations.

Furthermore, even at short wavelengths where Equations 19–16 and 19–17 suggest that one horizontal plane will suffice, it will sometimes be necessary to use two. The reason for this is simple, and has a direct analogy with deconvolution in two dimensions. Because the tangent plane separates from the "celestial sphere" away from the tangent point, the three-dimensional image lies below the tangent plane everywhere except at the tangent point. The image computed on the tangent plane alone is a slightly distorted form of this image. Two-dimensional deconvolution of this image leads to errors, since the beam is really three-dimensional and its structure in the n-direction varies from plane to plane. Precise deconvolution requires two (or more) image planes, and three (or more) dirty beam planes. The error introduced into the deconvolved image will scale roughly as the typical gradient of the sidelobe pattern multiplied by the separation between the unit sphere and the tangent plane.[3]

The analogy with two-dimensional deconvolution is straightforward. If a point of emission lies between cells in an image, it is well known that clean components must be used from many adjacent cells to correctly deconvolve the image. If we were constrained to subtract only from the cell closest to the peak of the emission, it should be easy to see that an enhanced error would result. Exactly the same situation occurs in three-dimensional image cubes. Deconvolving only on the tangent plane will always result in an error that grows quadratically with distance from the tangent point. For high dynamic range images, a multiple-plane deconvolution must be used, even if a simple analysis suggests that one plane is required.

The three-dimensional transform method of recovering the correct sky emission is attractive because it is conceptually simple and analytically precise. Tests by Tim Cornwell at the NRAO have shown that it correctly removes background sources from image cubes taken from VLA **D** and **C** configuration data at 327 MHz (see the examples in Sec. 3 of this lecture).

2.2. Polyhedron imaging

In this approach, one repeatedly applies the small field approximation, so that the "celestial sphere" is approximated by pieces of many smaller tangent planes. For each tangent plane, the visibility data must be phase shifted so the center of the new sub-image lies at the desired point on the sphere, and the (u, v, w) baseline components must be rotated so the resulting image is tangent to the sphere. These operations are illustrated in Figure 19–7.

The tangent plane represents the image formed by regular 2–D imaging. As I have argued, if the separation of planes on the n-axis exceeds half the synthesized beamwidth, unrecoverable errors will result. (This is a result of the sampling theorem—the separation of planes on the n-axis cannot exceed half the inverse baseline in wavelengths.) Figure 19–8 shows that the separation

[3] As an example, consider observing the well-known radio galaxy Cygnus A at 6 cm with the full resolution of the VLA (0.35 arcseconds). Its angular size (2 arcminutes) is such that the hot spots located at the extrema of the source are only 1/20 of a cell beneath the tangent plane. Assuming a peak-to-peak sidelobe level in the dirty beam of 1%, the typical slope, per cell, will be approximately 0.005. With an offset of 1/20 cell, the typical noise left by deconvolution due to misrepresenting the sidelobes will be approximately $1/200 \times 1/20 = 1/4000$, limiting the dynamic range to about 4000:1, consistent with that achieved in practice.

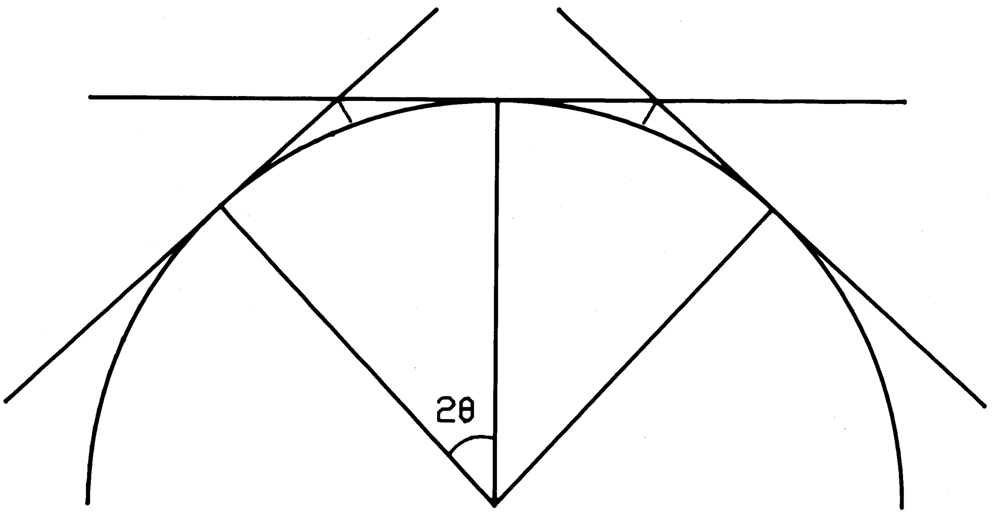


Figure 19–7. In polyhedron imaging, the image sphere is approximated by a polyhedron. The image on each polyhedral facet is produced by two-dimensional Fourier transformation of the data, suitably phase-shifted and rotated.

between the tangent plane and the image sphere is $1 - \cos\theta \approx \theta^2/2$, where θ is the angle from the phase-tracking center. Equating these gives an estimate of the *maximum* undistorted field-of-view in a two-dimensional image:

$$\theta_{\rm max} = \sqrt{\lambda/B} \approx \sqrt{\theta_{\rm syn}}\,, \tag{19–18}$$

where $\theta_{\rm syn}$ is the synthesized beam. Note that this is an unduly optimistic expression—as I explained above, being constrained to 'CLEAN' on one plane with good dynamic range reduces the allowable separation on the n-axis by perhaps a factor of 20, reducing the permitted field-of-view by a factor of 4 to 5.

Now consider the situation created by simply phase-shifting the image plane on the n-axis but not rotating the baselines, as shown in Figure 19–8.[4] If the new imaging center is denoted by θ_0, the separation between the image plane and sphere at offset θ is $\epsilon_2 = \cos(\theta_0) - \cos(\theta_0 + \theta) \approx \theta\theta_0$. Applying the critical sampling condition and assuming small angles, we find that

$$\theta_{\rm max} = \lambda/(2B\theta_0)\,. \tag{19–19}$$

This limit is much more severe than the tangent-plane limit. At the edge of the field-of-view, given by $\theta_0 \approx \lambda/D$, the "undistorted field-of-view" becomes $\theta_{\rm max} \approx D/2B$.[5] Using a more stringent condition for the allowable separation on the n-axis will reduce this angle by up to a factor of 5. The practical meaning

[4] The AIPS multi-field imaging and deconvolution program 'MX' operates in this approximation, to minimize computing time and to avoid re-sorting the data.

[5] For the VLA in its **A** configuration, this quantity is only $70''$, independent of frequency.

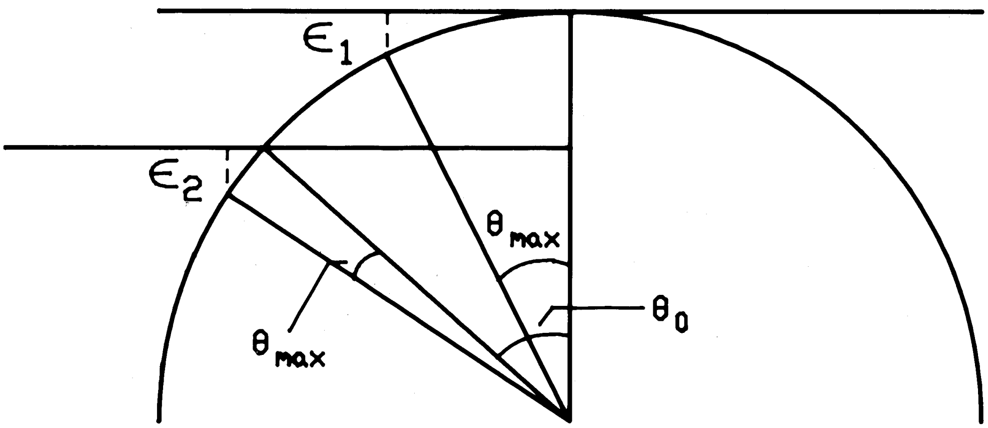


Figure 19–8. An illustration of the reduction of the undistorted field-of-view which is caused by not rotating the sub-image data. If the data are merely phase-shifted to the sub-image coordinate without baseline rotation, the image plane is shifted to the location of the desired new phase-tracking center. The new image plane remains parallel to the tangent plane. The region of undistorted imaging is given by the separation between the image plane and the image sphere. Not rotating the image plane greatly reduces the distortion-free region.

of this analysis is that not rotating the baseline components greatly reduces the "undistorted" field-of-view, and requires that many more images be made if it is important that all of the outlying fields be imaged with high fidelity.

The obvious question in the polyhedron imaging approach is: how many polyhedral facets are required? The required number of facets in the polyhedron can be estimated by the following elementary argument. Figure 19–7 shows the "celestial sphere" approximated by small, flat segments of tangent planes. To reduce errors, the maximum separation between tangent plane and sphere must be less than $f\lambda/2B$, where the factor f is a factor whose purpose I'll explain shortly. As shown, the actual separation between plane and sphere is $1 - \cos\theta$, and, applying the small-angle approximation, I find $\theta^2 < f\lambda/2B$. The number of images required to fill out the full primary beam is simply the ratio of the primary beam solid angle to the sub-field solid angle. Thus, the number of fields is

$$N_{\rm poly} = 2B\lambda/fD^2\,, \qquad (19\text{–}20)$$

where the symbols have the same meanings as in the last section.

The meaning of the factor f should now be clear. For critical sampling, $f = 1$, with the result that at least twice as many polyhedron images are needed as planes in the three-dimensional approach. But, as I argued in Section 2.1, high dynamic range imaging requires the separation between tangent plane and sphere to be much less than one.[6] This may seem impractical, but note that the size of the sub-fields is smaller, and considerable computation will be saved through this method. Indeed, it is easy to show that the full '3-d' inversion is

[6] For the case of Cygnus A with 0.35 arcsec resolution, $f = .05$, 20 times more sub-fields, or 40 times more images, are needed than in the three-dimensional approach.

completely impractical at low frequencies, so the 'polyhedron' method is the only simple alternative. The number of pixels in a full-beam 3-d image is roughly $4\lambda/D(B/D)^3$ – a number exceeding 400 million for **A**-configuration data at 327 MHz. However, the number of pixels required in the polyhedron method is approximately $8(B/D)^2$ – independent of wavelength. When the ratio, $\lambda B/2D^2$ greatly exceeds one (which it does for **A** configuration data at 21 cm, and all data at 90cm, except for the **D** configuration), it is advisable to employ the polyhedron method.

On the other hand, deconvolution by this method will be more expensive, since accurate removal of sidelobes will require subtraction from the ungridded data, an operation that is unnecessary in three-dimensional deconvolution. Furthermore, this technique requires the full database to be phase-shifted for each sub-field—another computational expense – for each subfield. Nevertheless, the 'polyhedron' method has been efficiently employed into the AIPS program 'IMAGR', with proper rotation of the baselines.

An alternate, but similar, approach may be attractive. Rather than reducing the size of the sub-fields to reduce deconvolution errors, one might consider computing two or three planes on the n-axis below the tangent plane for each rotated sub-image. Deconvolution could interpolate between these planes for a more accurate representation. A deconvolution that uses ungridded subtraction would again be needed, to remove the sidelobe structure correctly from adjacent sub-fields. The advantage is that the image sampling in the n-direction can be much larger, $f \approx 1$, so the number of computed images would be much reduced. Although this method should give superior results, it has not yet been tested.

2.3. Joint deconvolution

I remarked in the introduction that two-dimensional arrays do not, in general, generate coplanar visibility samples. A little thought should reveal that this statement is not true for short snapshots. If an array is sufficiently flat (more quantitatively, if the departure from coplanarity is less than some small fraction of the cell size in the n-direction, depending on dynamic range), each snapshot will allow a precise two-dimensional inversion. We might then contemplate the following approach:

1. Break the databases into a number of short observations. Within each, the array remains coplanar, as defined above.

2. Produce a two-dimensional image from each one.

3. Perform a joint deconvolution to output a deconvolved image.

The three items above deserve some comments. First, the length of the "short observation" in which the array remains coplanar is a function of what final dynamic range is desired. A scenario in which a strong background source located at the -10dB level of the primary beam will require much finer time divisions than one in which the confusing objects are only a few times the thermal noise.[7] Second, the geometry changes with each snapshot. That is, as the plane on which the visibility data are located rotates, the relation between the image coordinates (i.e., the cell locations) and the astronomical coordinates (α, δ) changes. Thus, as the sequence of images is made, the objects in the sky will appear to move along loci well removed from the positions given by the standard formulae. It can be shown that the cell positions on the snapshots are given by

$$\begin{aligned} l' &= l + \left(\sqrt{1-l^2-m^2} - 1\right) \tan Z \sin \chi \,, & (19\text{–}21) \\ m' &= m - \left(\sqrt{1-l^2-m^2} - 1\right) \tan Z \cos \chi \,, & (19\text{–}22) \end{aligned}$$

where (l, m) are given in Section 1.3, (l', m') are the cell coordinates of the snapshot image, and Z and χ are the zenith distance and parallactic angle at the time of the observations. For sources located much less than a radian from the phase tracking center, the positional differentials become: $\delta l = (r^2/2) \tan Z \sin \chi$, and $\delta m = -(r^2/2) \tan Z \cos \chi$, where $r = \sqrt{l^2 + m^2}$. These loci are shown in Figure 19–9. The parallactic angle and zenith distance are given by:

$$\begin{aligned} \tan \chi &= \frac{\cos \phi \sin H}{\sin \phi \cos \delta - \cos \phi \sin \delta \cos H} \,, & (19\text{–}23) \\ \cos Z &= \sin \phi \sin \delta + \cos \phi \cos \delta \cos H \,. & (19\text{–}24) \end{aligned}$$

There is no closed-form expression for the inverse of Equations 19-21 and 19-22; thus an iterative scheme is required to derive the standard direction cosines (l, m) from the given image coordinates. Third, note that the effect of this rotation of the (u, v) coordinates can be regarded as a non-uniform stretching of the image. The image wanders about the "standard" position, with the offset being given by $\Delta r = (\sqrt{1-l^2-m^2} - 1) \tan Z$, and the angle from the (l, m) axes given by $\tan \phi = -\cot \chi$. Figure 19–9 shows how the apparent position of an object moves as a function of declination and hour angle for the VLA. The offset is proportional to distance from the phase-tracking center, so it cannot be regarded as a shift. This non-uniform stretching precludes an approach based on "stretching" each snapshot image to a standard grid, summing the images, and then making one deconvolution. Unfortunately, nonlinear stretching to a standard grid destroys the convolution relation that holds for each snapshot individually. The only recourse is to deconvolve all the snapshots jointly. I am grateful to Fred Schwab for pointing out this problem, and its (painful) solution.

This approach would appear to have little to recommend it for imaging the entire primary beam at low frequencies. At high frequencies, a given array remains coplanar to the accuracy required by typical problems for long periods of time, since the primary beam is smaller and background sources are less troublesome. This approach may be useful for these situations. Again, it has not been tested, to the author's knowledge.

[7] As an example, consider a criterion of oversampling by a factor of 5, and imaging the entire primary beam (to 10 dB) at 327 MHz in the VLA's **A** configuration; I find that a new snapshot image must be made every minute! And, these images should have 8192 × 8192 cells!

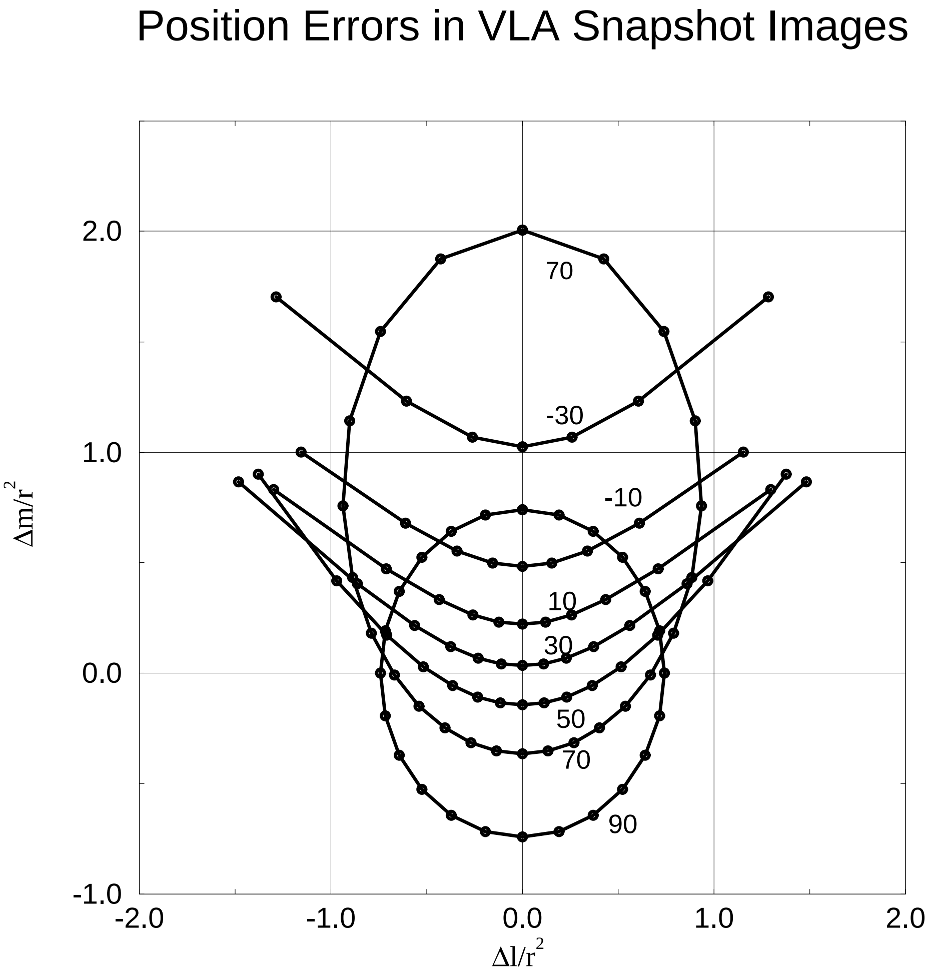


Figure 19–9. The positional error loci for an object observed with the VLA for seven representative declinations. The dots on the loci indicate the apparent position at hourly intervals, with the error at HA=0 extending along the vertical axis. All angular units are in radians.

3. Examples of Three-Dimensional Imaging

To test these ideas, trial software was written by Tim Cornwell on a Convex C-1 to do straightforward three-dimensional imaging and deconvolution. To test the concepts, he and I used VLA **C** configuration data for 3C 326 at 327 MHz, for which four planes are needed in the n-direction. Because this number is small, a three-dimensional FFT is unwarranted—aliasing in the n-dimension is bad. A direct, ungridded transform in this dimension was needed, so the image "cube" was built up as a series of two-dimensional transforms. For each plane, the phase of each visibility was rotated by $e^{2\pi i w N \delta n}$ before gridding, where N is the number of planes.

Figures 19–10 and 19–11 show some examples. The former shows five of the eight planes required in the beam, contoured in the standard fashion. (The

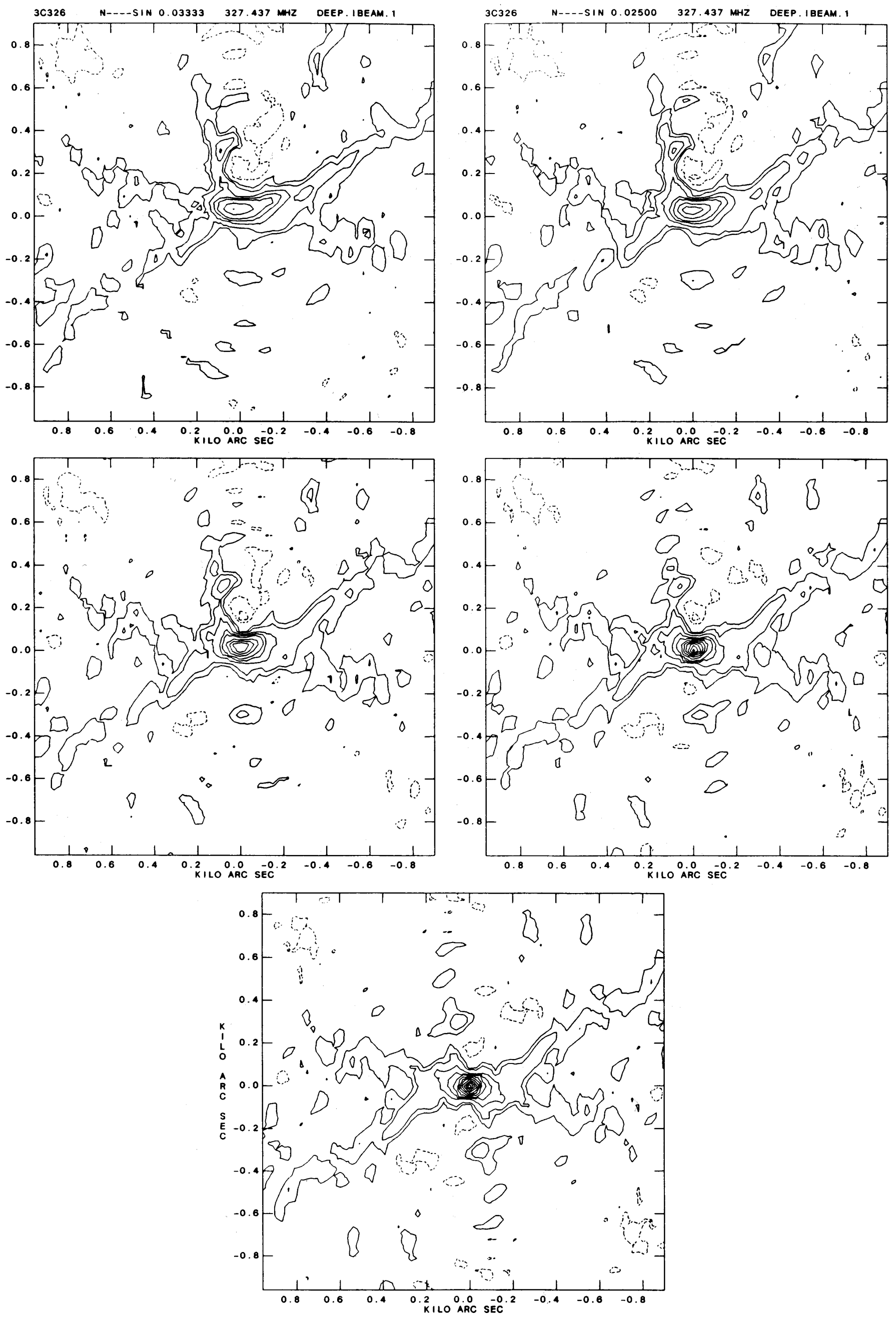


Figure 19–10. An illustration of the change in structure of the three-dimensional dirty beam. Five of the planes required to sample the beam in the n-direction for **C** configuration observing with the VLA at 327 MHz are shown. The bottom plane is in the upper left, while the central plane is at the bottom. The upper planes are identical to the lower, after rotation by 180°. The response to a point source in a two-dimensional image will degrade similarly with radius from the phase-tracking center.

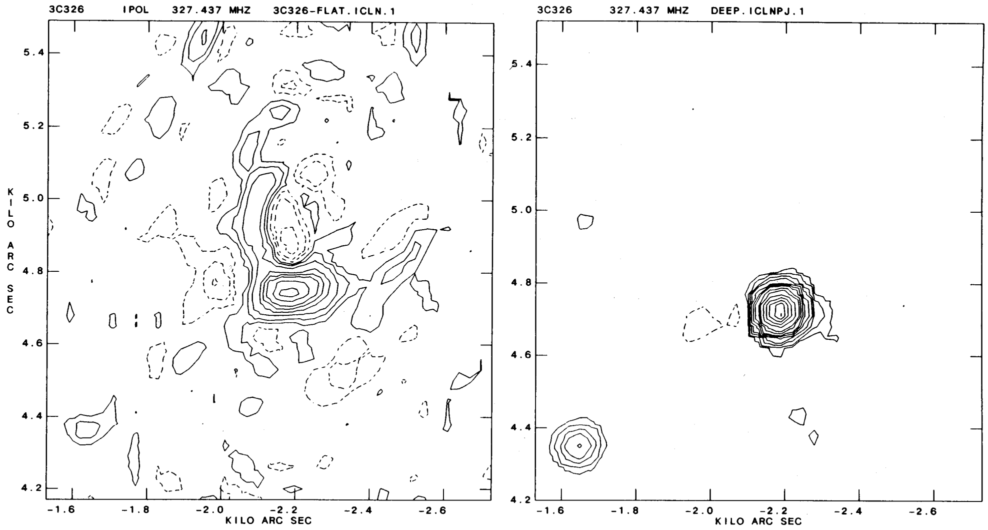


Figure 19–11. An example showing the results of deconvolving a tangent-plane image *(left)*, and a three-dimensional image *(right)*. Both are contoured at identical levels at 40 mJy per beam and above, but that on the right has an extra contour at 20 mJy per beam. The "flat" image is highly distorted, but the three-dimensional image appears to have no remaining distortions.

beam must always have twice as many planes as the image, although this can be cut in half if its symmetry properties are used.) The distinct broadening is caused by the use of natural weighting. Note that the beam is both broader and weaker on the lowest plane (upper left) than on the tangent plane (bottom). The response to a point source on a two-dimensional image will degrade in the same way as this beam with increasing distance from the phase-tracking center.

Figure 19–11 shows the 'CLEAN' images resulting from two-dimensional deconvolution (left) and three-dimensional deconvolution (right), followed by projection back onto the tangent plane for a strong background object located approximately 1°.5 from the phase tracking center. For the VLA's **C** configuration, the maximum response for this object is on the third plane. The result of two-dimensional deconvolution can be explained by reference back to Figure 19–10. The tangent-plane representation of this object resembles the upper right box of this figure, while the beam used to deconvolve it is in the bottom box. The **X**-shaped sidelobes are approximately the same, so these have been removed fairly well. However, the vertically rising spur in the dirty image is not in the beam (nor is the adjacent negative sidelobe), so the deconvolution program has no choice but to consider these features as real. The more exact deconvolution has made a striking improvement.

4. Acknowledgments

I thank Robert Braun, Barry Clark, Peter Napier and, especially, Tim Cornwell, for useful discussions on three-dimensional imaging.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

20. Mosaicing with Interferometric Arrays

M.A. Holdaway
National Radio Astronomy Observatory, Tucson, AZ 85721, U.S.A.

Abstract. Mosaicing is an image reconstruction technique by which multiple images are pieced together to form a single image. In radio interferometry, mosaicing refers to imaging objects larger than the primary beam, which therefore require multiple pointings on the sky. Because any object larger than the primary beam will also suffer from short spacing problems, radio astronomical mosaicing usually implies the addition of total power to the interferometric data.

1. Introduction

A given VLA configuration samples a limited range of spatial frequencies: the maximum baseline is about 40 times longer than the shortest baseline. This means that for a single configuration, single pointing image, you will have something like 40 resolution elements across an adequately sampled source. Our eyes see thousands of resolution elements across any field of view, so single configuration, single pointing radio images are usually not visually beautiful (though they can be intellectually very beautiful if they demonstrate a scientific point very clearly). In order to make a visually beautiful VLA image with many resolution elements across the source, you can observe in multiple VLA configurations (i.e., decrease the size of the resolution element) or perform a multiple pointing observation, or mosaic (i.e., increase the size of the source). Of course, your source has to cooperate by being bright enough at high resolution to make multiconfiguration observations, or by being large enough to warrant a mosaic.

Historically, there has been a major division in radio astronomy between single dish observations and interferometric observations. It was perceived that single dishes were required to image very large objects, and that only compact objects were suited to imaging with interferometers. However, with sensitive telescopes and modern mosaicing techniques combining total power and interferometric data, large sources have been observed at high resolution with fantastic results. For example, take a look at Figure 20–1. This is a mosaic of 58 VLA D array pointings at 1.4 GHz, with total power from one of the BONN surveys.

1.1. Large Sources Cause Problems

There are two criteria for “large source”:

- **Large compared to the reciprocal size of the shortest baseline,** $\theta > \lambda/b_{\rm min}$. Source structure larger than about $\lambda/b_{\rm min}$ is not reproduced in reconstructed images. Some sort of shorter spacing data must be measured. That might be interferometric data from a more compact configuration or from an array with smaller, closer dishes, or it might be total power data from a single dish.
- **Large compared to the primary beam,** $\theta > \lambda/D$**.** Imaging to the edge of the primary beam will effectively add noise to the outer parts of a source. Source structure beyond the primary beam is completely lost, of course. Mosaicing can answer both of these problems.

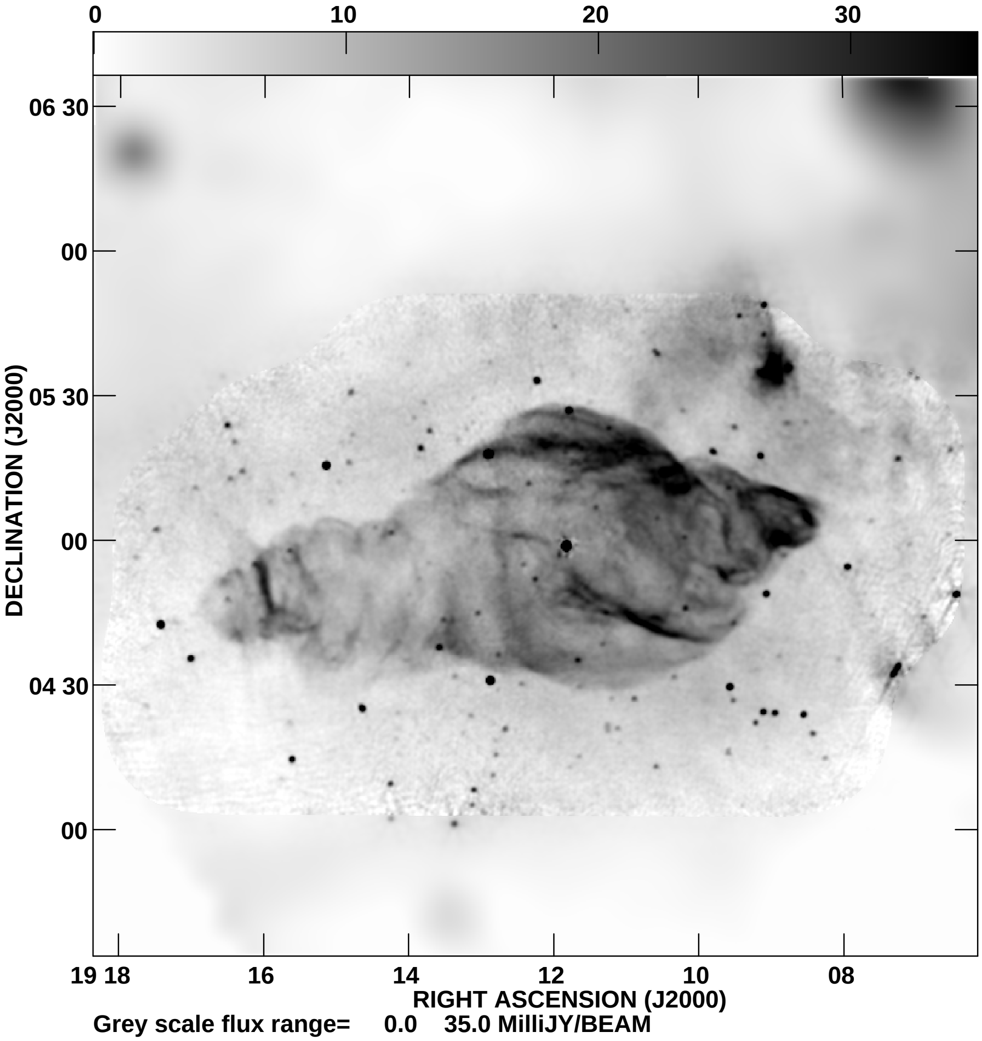


Figure 20–1. A radio continuum mosaic of the W50 SNR, from Dubner *et al.*

These two "large source" criteria are related in that $b_{\rm min} \geq D$, or else shadowing, or worse, colliding, is occurring. Fortunately, mosaicing algorithms can address both problems.

1.2. Single Pointing Imaging of Large Sources

Most interferometric imaging (to date) is concerned with only a single pointing on the sky. To demonstrate that a single pointing on the sky is not always satisfactory, we have simulated some VLA data on a source which is larger than the primary beam. Figure 20–2 a shows the model brightness distribution, gridded on 2.5 arcsec pixels. Figure 20–2 b shows the raw model brightness distribution smoothed with a 6 arcsec Gaussian beam, the approximate resolution of the

VLA's D configuration at 15 GHz. This is the image we would *like* to image with the VLA's D configuration. Figure 20–2 c shows a model of the VLA's primary beam at 15 GHz. Note the primary beam's first sidelobe at the edge of the image. And Figure 20–2 d shows the raw model multiplied by the primary beam, and smoothed with the 6 arcsec Gaussian beam. The measurement equation for an interferometer is

$$V(\mathbf{u}) \;=\; \iint \left(\mathcal{A}(\mathbf{x}-\mathbf{x}_p) I(\mathbf{x})\right) e^{-2\pi i(\mathbf{u}\cdot\mathbf{x})}\, d\mathbf{x}, \qquad (20\text{–}1)$$

(where $\mathbf{x}$ and $\mathbf{u}$ are 2-d vectors representing the sky and Fourier plane coordinates, and $\mathbf{x}_p$ is the pointing center) so the visibilities will be the Fourier transform of Figure 20–2 d, and will never know about anything outside of the primary beam, except for monstrously bright sources leaking through the low level sidelobes of the primary beam. Hence, Figure 20–2 d is the best we can *hope* to reconstruct for a single pointing observation.

Using model Figure 20–2 a, we simulate data by multiplying by the primary beam, inverse Fourier transforming, and degridding onto (u, v) coordinates calculated for the VLA D array every 60 s, for a range of ±3 hours from transit. Thermal (i.e., Gaussian) noise was added to the visibilities to emphasize certain results of the primary beam correction. Figure 20–3 a shows the image generated from a simulated single pointing observation. The simulated data were gridded, Fourier transformed, deconvolved using MEM (which is superior to CLEAN for extended sources such as our model), and smoothed with the clean beam. This image is an estimate of the quantity $\mathcal{A}(\mathbf{x}-\mathbf{x}_p)I(\mathbf{x})$, so we must divide by the primary beam image $\mathcal{A}(\mathbf{x})$ in order to get an estimate of the true sky brightness. To avoid dividing by zero, we blank the image wherever the primary beam is less than 0.1 of the peak. The single pointing image which has been corrected for the primary beam is shown in Figure 20–3 b. A far cry from what we would have liked and what we had hoped for.

Admittedly, these single pointing images are not very pleasing. There are three obvious problems:

- We are missing most of the source, due to the primary beam.
- We do not adequately reconstruct the part of the image which is within the primary beam; there is no extended structure around the bright dots, rather there is actually a "bowl", the famous interferometer image artifact caused when trying to image sources larger than the reciprocal of the shortest measured baselines. Indeed, even the compact sources do not have the expected brightness.
- Dividing by the primary beam to raise the source structure up to its expected brightness causes a non-uniform noise distribution in the image. This effect is clearly seen, with the noise a factor of 10 times larger where we cut off the primary beam.

1.3. Mosaicing Large Sources

The obvious next step to try is to point the antennas at several pointing centers, one after the other. So, we simulated some more data, again over ± 3 hours from

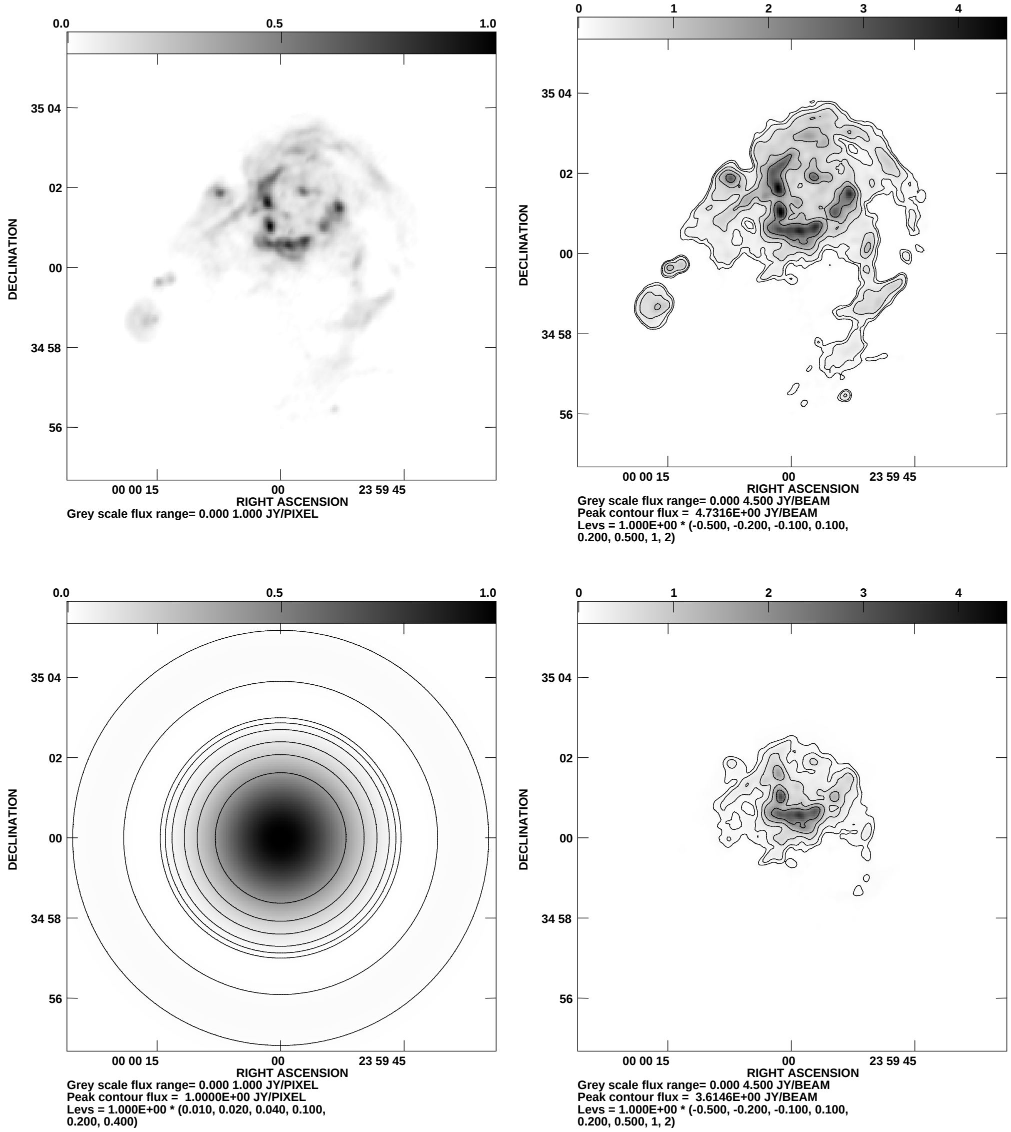


Figure 20–2. Model images for simulations: *a, top left:* Raw model brightness distribution. *b, top right:* Model brightness distribution smoothed with the clean beam. *c, bottom left:* Primary beam used for the simulations. *d, bottom right:* Primary beam times the raw model brightness distribution, then smoothed with the clean beam.

transit, but observing each of 9 pointings in turn. Now, when mosaicing, we need to get good (u, v) coverage for each pointing. Since the VLA's snapshot coverage is rather poor, it is advantageous to make very short snapshots of all pointings, and repeat over all pointings until you run out of observing time. So, instead of spending 6 hours on one pointing and getting good (u, v) coverage, we spend 1 minute on each of the 9 pointings, and repeat about 40 times to get a total

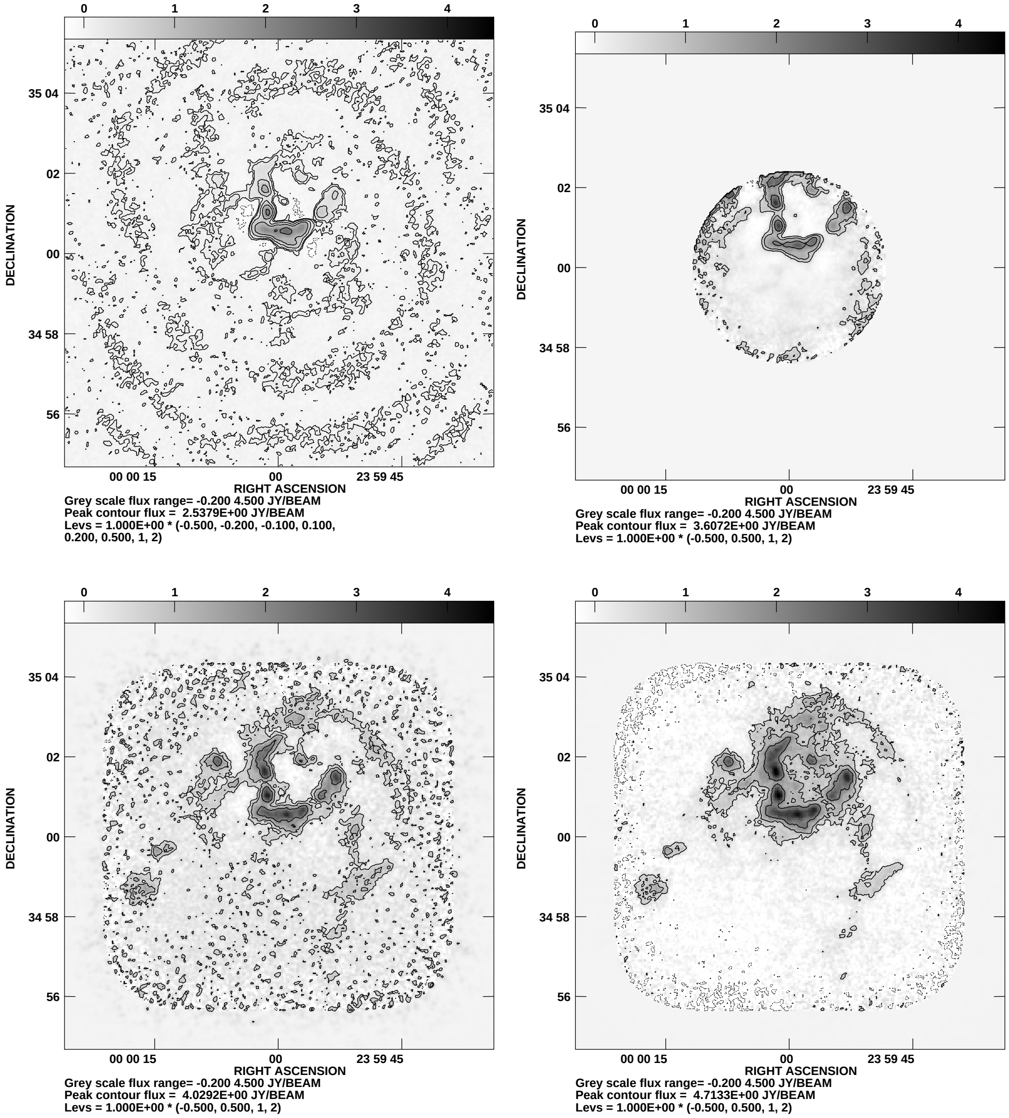


Figure 20–3. Model images for simulations: *a, top left:* Simulated VLA image of a single pointing. *b, top right:* Single VLA pointing after correction for the primary beam. *c, bottom left:* Nine VLA pointings deconvolved via a nonlinear mosaicing algorithm, without total power. *d, bottom right:* Nine VLA pointings deconvolved via a nonlinear mosaicing algorithm, including total power

of 40 minutes on each of 9 pointings, and still get good (u, v) coverage on each pointing. (In practice, some time will be lost between sources, but for mosaicing, good (u, v) coverage is more important than optimizing the sensitivity.)

The next problem is how do we convert the visibility data from these 9 pointings into one image? The first answer to this problem was to make 9 individual images, deconvolve them separately, and form a linear combination of the 9 images, considering the primary beam, as

$$I(\mathbf{x}) = \frac{\sum_{\mathrm{p}} \mathcal{A}(\mathbf{x}-\mathbf{x}_p) I_{\mathrm{p}}(\mathbf{x})}{\sum_{\mathrm{p}} \mathcal{A}^2(\mathbf{x}-\mathbf{x}_p)}. \tag{20–2}$$

Here, the summation is over pointing centers p, and $I_{\mathrm{p}}(\mathbf{x})$ is the p^{th} deconvolved image. This is the key mosaicing equation. If we have only a single pointing, it reduces to

$$I(\mathbf{x}) = I_{\mathrm{p}}(\mathbf{x})/\mathcal{A}(\mathbf{x}-\mathbf{x}_p), \tag{20–3}$$

which is just dividing by the primary beam, the inverse of Equation 20–1.

We don't show an image resulting from a linear combination of independently deconvolved images. In 1985, Tim Cornwell showed us that one can do much better by using all the data together to make a single image through a joint deconvolution. More on that later, but keep in mind that this is not just Fourier transforming all the data from all the different pointings; each pointings' data is Fourier transformed independently, and each pointings' primary beam is handled, and then one image which is consistent with all of that is formed. The result of a non-linear mosaic algorithm, a joint deconvolution using MEM, is shown in Figure 20–3 c. Now, we have come a long way in solving our problems, but there are still a few outstanding ones:

- Using enough pointings pointed at the right places on the sky allows us to catch the entire source.

- Performing a joint deconvolution actually improves the imaging of extended emission via the famous Ekers and Rots (1979) scheme (more on that later). The source still sits in a bowl, but it isn't as bad. A slightly smaller source might be imaged quite well by mosaicing only the interferometer data.

- Note that the noise has increased over the single pointing image. We spend 1/9 as much time per pointing now, so the noise will be $\sqrt{9} = 3$ times higher for an individual pointing. However, the increase in overall noise is augmented by a 1.4 factor decrease due to overlapping pointings (i.e., several neighboring pointings contribute to the imaging of a given pixel). The noise still flares up at the image edges, but the edges are now beyond the region of interest. Sault *et al.* (1996) have an alternative approach in which the noise does not flare up, but the flux scale varies at the mosaic edges.

What do we need to do to make a good image of our model source? We need total power. The image still suffers because we've got a hole in the center of our Fourier plane sampling, and the source we are trying to image changes significantly in the region we aren't sampling. How do we sample that region? The easiest way is to use a single dish. If a single dish just stares at one point on the sky, it is only measuring the $(0, 0)$ Fourier sample of the sky brightness times the single dish primary beam. However, if you scan across a source to image it and Fourier transform the image, you find that you have information all the way out to the dish diameter D. So, we need a single dish which is as big as the hole in the center of the Fourier plane.

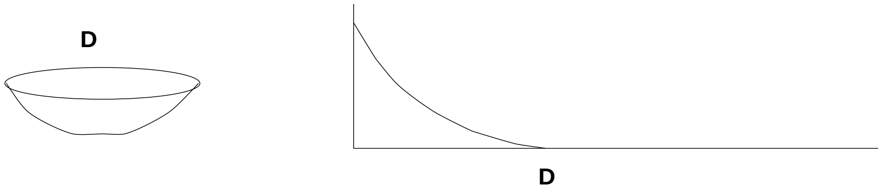


Figure 20–4. A single dish of diameter D and its distribution of "Fourier coverage".

There are several ways to incorporate total power into an image; we'll discuss some of them later. This simulated total power data was actually measured by the same dishes as are measuring the interferometric data in the homogeneous array scheme which will be used by the Millimeter Array (MMA) (Cornwell, Holdaway, and Uson, 1993). Since the total power is observed at the same pointings as the interferometer data, and since both have the same primary beam, it is possible to simply grid the total power as the $(0,0)$ point in each of the pointings' Fourier plane. That is the only thing which is different between Fig. 20–3 d and Fig. 20–3 c; otherwise, the processing is identical. With total power added, the extended emission is finally accurately imaged.

2. Ekers and Rots and Effective (u, v) Coverage for Mosaicing

Now that we've demonstrated what mosaicing can do for you, we'll take a look at how it works. The glue that holds mosaicing together was invented by Ekers and Rots (1979), long before mosaicing was truly invented. Ekers and Rots argued that each interferometric baseline measures not one spatial frequency, but a range of spatial frequencies.

Lets start with a single dish. As we already said, a single dish doesn't measure a single spatial frequency, but a range of spatial frequencies out to the maximum baseline. This is shown graphically in Figure 20–4. This curve, or the distribution of Fourier coverage, represents a 1-D slice through the density of various spacings measured by a single dish. You may say that a single dish doesn't *have* baselines. But you can think of the single dish as being made up of hundreds of little tiny patches, each like a an element of an interferometer. They are correlated physically when their signals are combined at the focus. Now, there are a great many of the shortest kinds of spacings between the little patches, fewer intermediate length spacings (as the dish starts crowding), and almost no spacings which span the entire dish diameter.

Ekers and Rots extend this reasoning to an interferometer of baseline b. A single visibility will be a linear combination of the visibilities obtained from all possible combinations of patches on one antenna and patches on the other antenna, with the shortest spacings being $b - D$ and the longest spacing being $b + D$. Figure 20–5 shows this graphically, along with the total power response from the two dishes. But how do we decode the information about all of the different spacings from our single visibility? You can't solve for N unknowns (i.e., Fourier information at several points between $b{-}D$ and $b{+}D$) with only one piece of data (i.e., one measured visibility); so, as the astronomical adage goes, "we

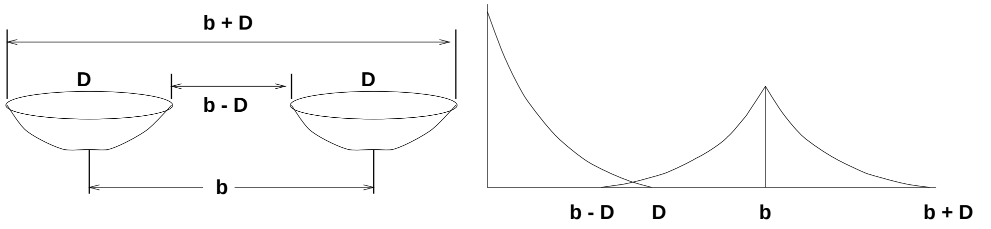


Figure 20–5. Distribution of the "Fourier coverage" of an interferometer with baseline b, including total power. The (u, v) coverage from 0 to D is from the two single dishes, and the (u, v) coverage between $b - D$ and $b + D$ is due to the interferometric data.

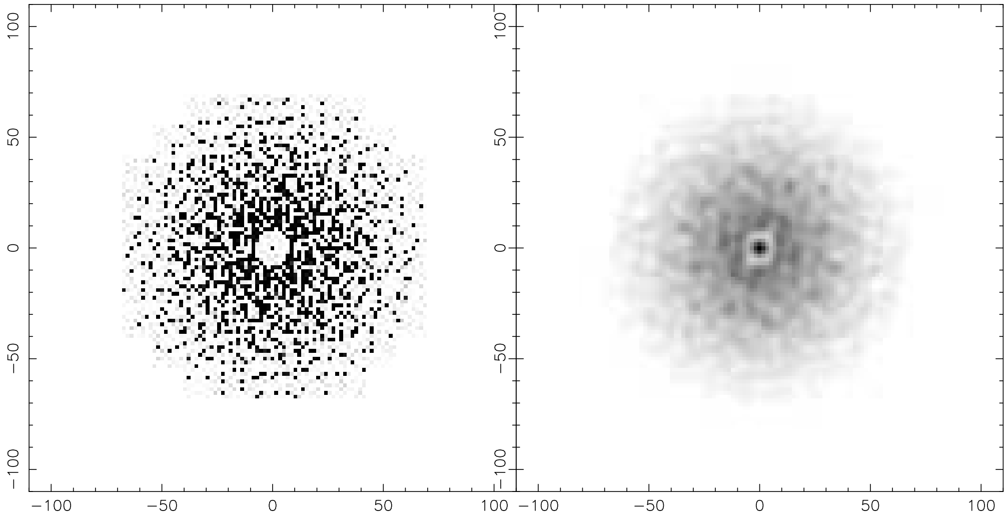


Figure 20–6. The effective (u, v) coverage of a sample compact MMA configuration.

need more data". In the single dish case, we were able to obtain the information about other spacings by scanning over the source and Fourier transforming. Ekers and Rots obtained the information between spacings $b - D$ and $b + D$ by scanning the interferometer over the source and Fourier transforming the single baseline visibility with respect to the pointing position. Another way of thinking about this is that changing the pointing position on the sky is equivalent to introducing a phase gradient in the (u, v) plane, which applies a different set of complex weights to the visibility information over the range of baselines. Each extra pointing we make provides a different set of weights, or provides a different equation that enables us to solve for more of those N unknowns, or allows us to extract the extended Fourier information from that one baseline.

A more mathematically precise statement of Ekers and Rots is that each sample in the Fourier plane should be replaced by the autocorrelation of the dish's illumination pattern. This is demonstrated strikingly in Figure 20–6, which shows the snapshot Fourier plane coverage for a candidate for the MMA's compact configuration as delta functions (slightly magnified in width for your viewing pleasure), and also the effective Fourier plane coverage for mosaicing, which has been convolved with the autocorrelation of the dish's illumination pattern. We have just applied the Ekers and Rots scheme to the entire array! While the delta function Fourier plane coverage has lots of holes, the effective Fourier plane coverage is essentially complete. This is a good thing. A 2-D imaging interferometer will generally not have complete Fourier plane coverage, and generally doesn't need it; most radio sources are fairly discrete, taking up only a small part of the primary beam. Using clean boxes to ignore regions with no detectable radio emission results in the situation of having more unique Fourier samples than the number of pixels we are trying to solve for, thereby permitting good quality images. When we are mosaicing, every field observed is potentially filled with complicated source structure, and it is not possible to reduce the number of imaged pixels with clean boxes. In order to make good images, we *need* complete effective (u, v) coverage. With instruments like the VLA, it takes a dozen independent snapshots well distributed in hour angle to get pretty good effective (u, v) coverage, which severely hampers mosaicing. And its even worse for a 1-D array like WSRT or the AT. Any instrument that will be doing a lot of mosaicing should strive for complete effective (u, v) coverage in a single snapshot.

If you've been paying attention, you've probably noticed something odd in the center of the effective (u, v) image that looks like an eye, a dark pupil surrounded by a light colored iris. A slice through the effective (u, v) coverage of the homogeneous MMA array is shown in Figure 20–7. The dark central pupil is due to the total power data, measured by the same dishes that measure the visibilities, but auto-correlated instead of cross-correlated. As each of the many dishes is contributing total power measurements, the sensitivity at the center is rather large. The lighter iris represents a decrease in the effective (u, v) density at the spacings where the total power sensitivity is decreasing, but the shortest interferometric baselines haven't yet kicked in very strongly, roughly the region between $b - D$ and D in Figure 20–5. If we had used a single dish of diameter larger than the interferometric dishes, the total power sensitivity would have extended further out in Fourier space, largely eliminating the sensitivity dip. However, a single feed large single dish would have had much lower sensitivity and would have required multi-feed arrays in order to be comparable to the shortest baselines. Furthermore, total power measurements are much more difficult to make than interferometric measurements and are more susceptible to systematic errors. In spite of the sensitivity dip, simulations indicate that the homogeneous array (i.e., equipping all antennas in the interferometric array with total power systems) can make excellent quality images and may even be superior to using a large single dish for the total power measurements. A slice through the effective (u, v) coverage of the homogeneous MMA array is shown in Figure 20–7.

3. Mosaicing Algorithms and Total Power

Now that we've seen some of the theoretical underpinnings that make mosaicing work, we'll take a look at the algorithms people have implemented to perform mosaicing in practice.

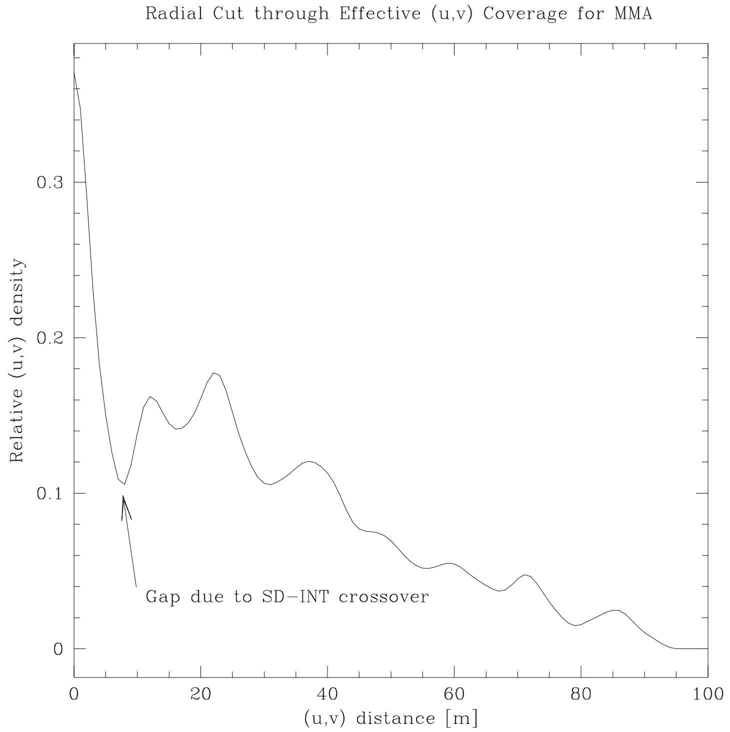


Figure 20–7. Slice through the effective (u, v) coverage of a sample compact MMA configuration.

3.1. Mosaicing Previously Deconvolved Images

As mentioned above, we could take the data from each pointing center and Fourier transform and deconvolved each separately, and then mosaic the separate images using Equation 20–2. This method does not exploit the Ekers and Rots scheme as only one sky pointing is handled at a time, so there is no way to separate out the range of spatial frequencies potentially available from each baseline.

In this algorithm, total power data was added to the interferometer data for each pointing in a somewhat circuitous manner: first, the total power data was obtained with a dish of diameter much larger than the central hole to be filled. Next, the single dish data was deconvolved to remove the effects of its beam. The deconvolution is poorly determined at the spatial frequencies at the dish edge, which is why a very large dish is favored, so the spatial frequencies being used are well determined. The interferometer's primary beam for a given pointing is applied to the deconvolved total power image, and fake interferometer tracks of very short baselines are "simulated" from this image. The fake baselines are chosen so as to fill in the hole in the interferometer's Fourier plane. In hardware, you can't have a baseline shorter than the dish diameter without shadowing. In software, there is no such constraint. Simulations of this method indicate that it is usually vastly inferior to joint deconvolution methods.

3.2. Non-linear Joint Deconvolution

The first approach to mosaicing which incorporated Ekers and Rots information was made by Cornwell (1985; 1988). I don't know of any algorithm which explicitly uses the Ekers and Rots information by taking all the visibilities from a given baseline, gridding them as to sky pointing position, and Fourier transforming them to obtain the range of Fourier plane information actually sampled. Rather, algorithms which perform a deconvolution using information from all the pointings at once to generate a mosaiced image utilize all the Ekers and Rots information present in the effective (u, v) coverage implicitly. This can be demonstrated in simulations by comparing the mosaiced image with the true brightness distribution in the Fourier plane.

The non-linear joint deconvolution mosaicing algorithm basically finds an image which is consistent with all the measured data, or which optimizes a global (i.e., formed from all the pointings' data) χ^2, which is a measure of the goodness of fit. Cornwell defined the global χ^2 as

$$\chi^2 = \sum_{\rm p}\sum_{\rm i} \frac{|V(\mathbf{u}_{\rm i},\mathbf{x}_{\rm p}) - \hat{V}(\mathbf{u}_{\rm i},\mathbf{x}_{\rm p})|^2}{\sigma^2_{\rm V}(\mathbf{u}_{\rm i},\mathbf{x}_{\rm p})}, \qquad (20\text{–}4)$$

where the sums are over all visibilities i in each pointing p, and $\hat{V}$ is the model visibility calculated according to Equation 20–2 from the current best model of the brightness distribution. The gradient of χ^2 with respect to the model image basically tells us how we can change the model image so that χ^2 is reduced (i.e., so that the image agrees more closely with the measured data from all pointings). The gradient can be used in a maximum entropy routine (or another optimization program) to solve for a model brightness distribution which is consistent with the noise. The form of the gradient image is instructive:

$$\nabla_{\hat{\rm I}}\chi^2(\mathbf{x}) \;= -2\sum_{\rm p}\mathcal{A}(\mathbf{x}-\mathbf{x}_p)\left(I^D_{\rm p}(\mathbf{x}) - B_{\rm p}(\mathbf{x}) * (\mathcal{A}(\mathbf{x}-\mathbf{x}_p)\hat{I}(\mathbf{x}))\right), \qquad (20\text{–}5)$$

where $\hat{I}$ is the model image, the sumation is over pointings p, $I^D_{\rm p}(\mathbf{x})$ is the dirty map for the p^{th} pointing, and $B_{\rm p}(\mathbf{x})$ is the point spread function for that pointing. While this looks ugly, it isn't difficult to understand. Basically, this is like the numerator of Equation 20–2. The gradient image is an unnormalized mosaic of the residual images for each pointing. Each residual image, which is the expression between the large parentheses, is formed by multiplying the global model image by the primary, convolving with the dirty beam, and subtracting that from the dirty image. As the model image comes closer to reality, the gradient image becomes smaller. If we were able to achieve a perfect reconstruction, the gradient image would be only noise. One of the shortcomings of the MEM-based mosaicing is that the residuals, or gradient image, usually has some structure in it upon convergence, indicating that the algorithm hasn't worked exactly as we had hoped. This is a topic of ongoing research.

This mosaic algorithm is non-linear because it relies upon MEM, a non-linear deconvolver. In contrast, we can refer to the linear combination of previously deconvolved images, as specified in Equation 20–2, as a linear mosaic. We refer to the current algorithm as a joint deconvolution because we are trying to deconvolve everything in one go.

Total power data can be incorporated into this mosaicing algorithm in a number of ways:

- As mentioned before, if the array is able to measure total power at the same point positions and with the same primary beam as the interferometer (which are met if we are able to measure total power with the interferometric elements at the same time we are measuring visibilities), we can simply include the total power data as the $(0,0)$ Fourier points for each pointing's data.

- If the total power data have the same primary beam as the visibilities, but were measured at different sky positions, they can be regridded onto the pointings the interferometer visited and treated as above.

- If the total power's primary beam is different from the interferometer's, or if the total power data were taken at different sky positions than the interferometer data, we can treat each total power data point as a separate pointing with its own data set of one number at the $(0,0)$ point, each contributing to the $\nabla\chi^2$ image. In order to contribute in a reasonable way, we might need to fiddle with the relative weights (i.e., $1/\sigma^2$) of the total power and interferometer data. (We have glossed over weighting in the mosaicing equations.) An example of mosaicing total power with interferometer data which uses this method is shown in Figure 20–8.

- When the total power is measured with a single dish that is much larger than the interferometer's elements, there will be considerable overlap between the longest single dish spacing and the shortest interferometer spacing. In this case, one can achieve good results by using the total power image, converted to units of Jy/pixel (i.e., deconvolved unit) as a bias image for the MEM-based mosaic program.

- One can mosaic the interferometric data using the MEM-based mosaic program and merge the total power image with the interferometer-only mosaic image in the Fourier plane with an algorithm such as IMERG. This is not recommended because the deconvolution without the benefit of the total power data will result in bowls and leave artifacts which will not be completely undone by adding the total power after the fact. Furthermore, edge effects can lead to painful problems in this method.

Non-linear joint deconvolution mosaic algorithms are implemented in SDE as `mosaic` and in AIPS as `vtess` and `utess`.

3.3. Linear Mosaic of Dirty Images with Subsequent Joint Deconvolution

For reasonable sized mosaics observed with an intelligent observing strategy of multiple short snapshots on each pointing, each pointing will have fairly similar (u, v) coverage and hence fairly similar point spread functions. Its a reasonably good approximation to assume that each pointing has an identical point spread

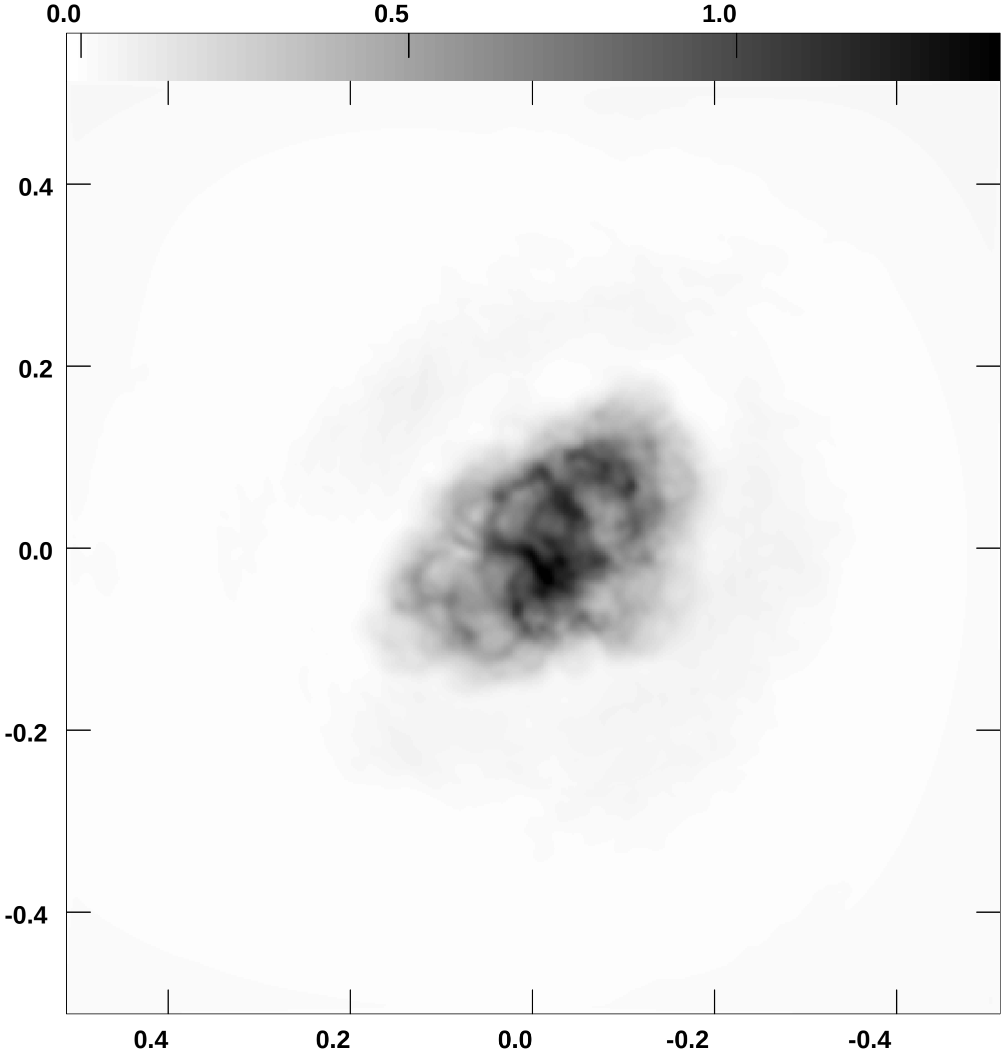


Figure 20–8. Mosaic image of the Crab nebula at 8.4 GHz from Cornwell, Holdaway, & Uson (1993). This is an example of a homogeneous array mosaic combining VLA data and total power from a VLBA antenna. The total power data was included as separate pointings in the non-linear mosaic algorithm.

function. Under this approximation, we are able to form a linear mosaic of the *dirty* images as specified by Equation 20–2, and deconvolve the single image with a slightly modified effective point spread function. In the case where the effective (u, v) coverage is complete, we can even deconvolve with a linear method such as Wiener filtering. However, CLEAN or MEM will work fine as well.

Now, this algorithm can produce images with dynamic ranges of a few hundred to one, limited by the differences in the point spread functions among the pointings. However, we can use a trick invented by Barry Clark to make this method work to higher dynamic ranges. After we've deconvolved to the point where we are starting to be limited by the differences in the point spread functions across the mosaic, we take the partially deconvolved model image and form the mosaiced residual image as given in Equation 20–5 using the exact PSF for each pointing, and proceed with joint deconvolution of the residual image. In doing so, all of the bright emission gets removed from the data, and the "few hundred to one" dynamic range limitation now applies to whats left in the residual mosaic image. In principle, we can keep going like this until we hit the thermal noise limit, unless there are systematic errors in the data or our understanding of the data, which will be addressed further on.

The linear mosaic algorithm is implemented in SDE as `moslin`, in AIPS++ as part of `sky`, and in classic AIPS as `ltess`. The AIPS++ linear mosaic algorithm uses the minor/major cycle trick. The `ltess` program does not construct an effective point spread function for deconvolving a linear mosaic of dirty images.

3.4. The Sault *et al.* Algorithm

Sault, Stavely-Smith, and Brouw (1996) have implemented another joint deconvolution mosaicing algorithm in Miriad as a three step process of imaging, deconvolution, and restoring. First, a linear combination of the dirty images from each pointing are linearly mosaiced almost according to Equation 20–2. Sault *et al.* use an image plane weighting function which basically results in constant thermal noise across the image. This weighting function removes the upturn in noise which occurs at the edge of the sensitivity pattern, but source structure which is at the edge of the sensitivity pattern is not imaged at full flux. Mosaic images weighted this way are easier to view in 3-D data cube displays and are generally more appealing to look at. If a guard band is observed around the source of interest, there will be no source structure affected by the variable flux scale across the image.

The point spread functions or dirty beams from each pointing are stored in a cube. The gradient of χ^2 is formed as suggested by Equation 20–5 and used in MEM and CLEAN based deconvolution algorithms.

Total power could be added in a seamless fashion in this mosaic algorithm for a homogeneous array. For the case where the total power is measured by an antenna of different diameter than the interferometer's antennas, Stanimirovic *et al.* (1998) suggest a different method: given a mosaic of dirty interferometer images $I^D_{\rm int}$ and dirty beam $B^D_{\rm int}$ and single dish image $I^D_{\rm sd}$ and beam $B^D_{\rm sd}$, composite dirty images and beams are formed as

$$I^D_{\rm comp} \quad = \quad (I^D_{\rm int} + \alpha I^D_{\rm sd})/(1+\alpha) \qquad (20\text{–}6)$$

$$B^D_{\mathrm{comp}} \quad = \quad (B^D_{\mathrm{int}} + \alpha B^D_{\mathrm{sd}})/(1+\alpha). \tag{20–7}$$

$$\tag{20–8}$$

Then, a convolution relationship exists between the true sky brightness distribution I and the composite dirty image and the composite dirty beam:

$$I^D_{\mathrm{comp}} = I * B^D_{\mathrm{comp}}, \tag{20–9}$$

and I can be solved for from the composite dirty images and beams with the standard deconvolution algorithms MEM or CLEAN. But does it work? Adding the single dish's beam to the interferometer's beam can be understood in the image plane as adding a very broad plateau and reducing the negative sidelobes of the point spread function, thereby reducing the inclination towards producing "bowls". In the Fourier plane, adding the single dish beam basically fills the hole in the center of the (u, v) plane, as is seen by Fourier transforming the composite dirty beam.

The method is only approximate and breaks down when the single dish beam is comparable to the primary beam of the interferometer. Formally, the method will work for any α. Stanimirovic *et al.* suggest using an α which is the ratio of the beam area of the interferometer's synthesized main beam and the beam area of the single dish's main beam. I would argue that it makes more sense to base α upon the ratio of the sensitivities of the single dish and total power instruments. Using a very small value for α results in very slow convergence of the deconvolution algorithms.

A spectacular example of the Sault *et al.* (1996) and Stanimirovic *et al.* (1998) algorithms in action can be seen in Figure 20–9, a series of mosaics of a single channel of HI data from the AT's observations of the Small Magellanic Cloud. Panel *a* shows the mosaic of the interferometer-only dirty images, panel *b* shows the deconvolved interferometer-only image, panel *c* shows the total power image, and panel *d* shows the deconvolution of the composite image with the composite beam.

The Sault *et al.* programs are implemented in the MIRIAD system as `invert` (to make the dirty images and beams and to perform a linear mosaic of the dirty images), `mosmem` (to perform a MEM deconvolution), and `mossdi` (to perform a Steer-Dewdney CLEAN deconvolution).

4. Mosaicing as a Design Tool for Interferometric Arrays

There are various rules of thumb floating around for the required pointing accuracy and surface accuracy for the antennas of an interferometric array. Obviously, as the frequency increases, the pointing as a fraction of the primary beam and the surface errors as a fraction of a wavelength degrade. The rules of thumb for single pointing interferometry of compact objects are derived primarily from sensitivity considerations: if our pointing is too far off too much of the time, we start to lose sensitivity, and if the surface errors are too large a fraction of the observing wavelength, then Ruze losses become appreciable. Mosaicing imposes much more severe requirements on the pointing and surface errors for the antennas of a mosaicing interferometric array such as the MMA.

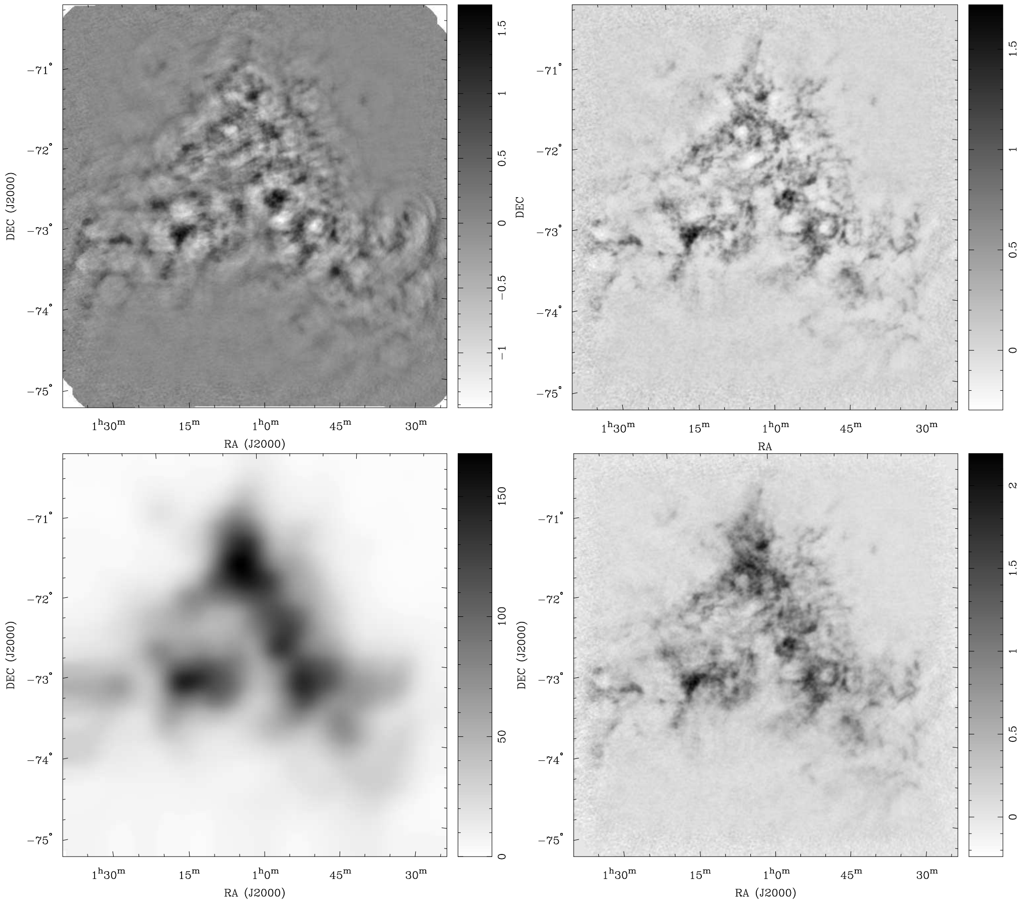


Figure 20–9. Processing of one channel of HI from the Stanimirovic *et al.*'s SMC mosaic. *a, top left:* a "dirty mosaic", interferometer data only. *b, top right:* a deconvolved mosaic, interferometer data only. *c, bottom left:* total power data from Parkes. *d, bottom right:* a deconvolved mosaic, including both total power and interferometer data.

Why are the requirements on the antennas tighter once we are considering mosaicing? The center of the primary beam is fairly flat, and if a small object is at the center of the primary beam, even large pointing jiggles such as 1/10 of a beam width will produce fairly small amplitude errors on the source. When mosaicing, we generally have emission everywhere within the primary beam. If there is some emission at the half power point, a small pointing jiggle will result in a substantial error in the amplitude of the part of visibility associated with that emission. While these errors are antenna based amplitude errors, they will not be fixed by common amplitude self-calibration: some emission on the opposite side of the primary beam will suffer the opposite amplitude error, and the effects of the pointing errors are dependent upon the source structure. Furthermore, the effects of the pointing errors depend highly upon how systematic or random they are in time and across the array.

Surface errors in the primary or secondary reflectors will result in primary beam sidelobes. Accurate mosaicing requires that the model of the primary beam used in the algorithm be a good match to the true primary beam. Systematic surface errors, as might be caused by systematic dish deformations due to gravity or solar induced thermal gradients on the antenna structure, could have a major effect on the mosaic quality. There will also be random surface errors due to the settings of the panels on the antenna backup structure, resulting in different voltage patterns for each antenna, ending in mosaicing errors which would limit the dynamic range of mosaic images.

Cornwell, Holdaway, and Uson (1993) provide a quasi-analytical treatment of the results of these errors on mosaic image quality. The effects of pointing errors are roughly proportional to the rms pointing error as a fraction of the beam, and the effects of surface errors are roughly proportional to the square of the rms surface error as a fraction of the observing wavelength. Hence, at low frequencies, pointing errors will limit mosaic image quality of very high SNR images, while surface errors will limit high SNR mosaic image quality at higher frequencies. The detailed understanding of the relative impact of systematic and random errors may be obtained by modeling the errors and performing numerical simulations of their effects on the visibilities and the final mosaic images.

5. Good Mosaicing Practice

- Point in the right place on the sky.
- Nyquist sample the interferometer and single dish pointings on the sky: pointings should be separated by $\lambda/2D$ if they lie on a rectangular grid.
- Observe extra pointings in a guard band around the source.
- Get total power. BONN's web site has total power continuum images at a few different frequencies. You may have to get total power data the old fashion way: observe.
- Observe many, many (i.e., 5-15) short snapshots on each pointing position to get very good (u, v) coverage, and very similar coverage across the source.
- Very large mosaics at the VLA can be sped up by using `//OF` cards. You lose 20 s at the VLA just by changing sources in the online system. However, using `//OF` cards (offset cards) in your observe file, you are able to move to new positions on the sky while fooling the on-line system into thinking that you've not changed sources. You will lose 3-6 s between each pointing position for move time. FILLM understands that these pointings are all different "sources".
- For calibrating and processing in AIPS, give careful thought to organization. Much effort can be saved by using wildcards, qualifiers, run files, and procedures.

References

Cornwell, T.J. MMA Memo No. 32, NRAO, 1985.

Cornwell, T.J. 1988, *A&A*, 202, 316-321.

Cornwell, T.J., Holdaway, M.A., & Uson, J.M. 1993, *A&A*, 271, 693–713.

Dubner, G.M., Holdaway, M.A., Goss, W.M., & Mirabel, I.F. 1998, *AJ*, 116, 1842.

Ekers, R.D., & Rots, A.H. 1979, in Proc. IAU Coll. 49, *Image Formation from Coherence Function in Astronomy*, C. van Schooneveld, Ed., D. Reidel (dorrecht. Holland), pp. 61–66.

Sault, R.J., Staveley-Smith, L. & Brouw, W.N. 1996, *A&AS*, 120, 375–384.

Stanimirovic, S., Staveley-Smith, L., Dickey, J.M., Sault, R.J. & Snowden, S. L. 1999, *MNRAS*, 302, 417.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

21. Multi-Frequency Synthesis

R.J. Sault
Australia Telescope National Facility, Epping, NSW 2121, Australia

J.E. Conway
Onsala Space Observatory, S-439 00 Onsala, Sweden

Abstract. Multi-frequency synthesis is the practice of using visibility data measured over a range of frequencies when forming a continuum image. Because observing frequency is easier to vary than antenna location, it is an effective way of filling the (u, v) plane for an observation. Here we consider the artifacts in MFS images caused by source spectral variation. For frequency ranges of about 30%, for observations where only modest dynamic range is required, the artifacts of MFS can be completely ignored. For higher dynamic range observations, some calibration techniques and deconvolution algorithms are described which minimize the artifacts.

1. Introduction

In imaging in interferometry, the (u, v) coordinate used in the imaging process is equal to the projected physical baseline length divided by the observing wavelength. That is, the (u, v) coordinate is measured in wavelengths. Thus, for a given physical antenna separation, we can measure different (u, v) coordinates by varying the observing wavelength (or frequency). Multi-frequency synthesis (MFS) is the practice of putting this characteristic to good use, and of using visibilities measured at a variety of frequencies in forming an image. The advantage of doing so is the better (u, v) coverage that can be achieved, and so the fidelity of the resultant image can be improved. For example for the VLBA, using 8 frequencies over a 15-30% frequency range, MFS can produce a dramatic improvement in (u, v) plane coverage (e.g. see Fig. 21–1).

Although this principle has been appreciated for some time (e.g. McCready et al. 1947), it has received little attention until the last ten years. This was probably a result of both hardware limitations and a lack of understanding of the artifacts created by the variation in source emission with frequency. The advent of wideband receivers and comparatively cheap and flexible digital correlators have made multi-frequency synthesis a much more attractive way of filling the (u, v) plane than building extra antennas. This is particularly so when the antennas are not reconfigurable into a variety of arrays (e.g. VLBI arrays). Recent analyses of frequency ranges up to about 30% have shown that the image artifacts introduced by multi-frequency synthesis are comparatively minor. Furthermore, calibration and imaging techniques have been developed to reduce these artifacts to substantially smaller levels.

This chapter considers ways of representing the effect of spectral variation in images formed from multi-frequency data. This naturally leads to calibration techniques and deconvolution algorithms that can be used to reduce significantly the resultant artifacts.

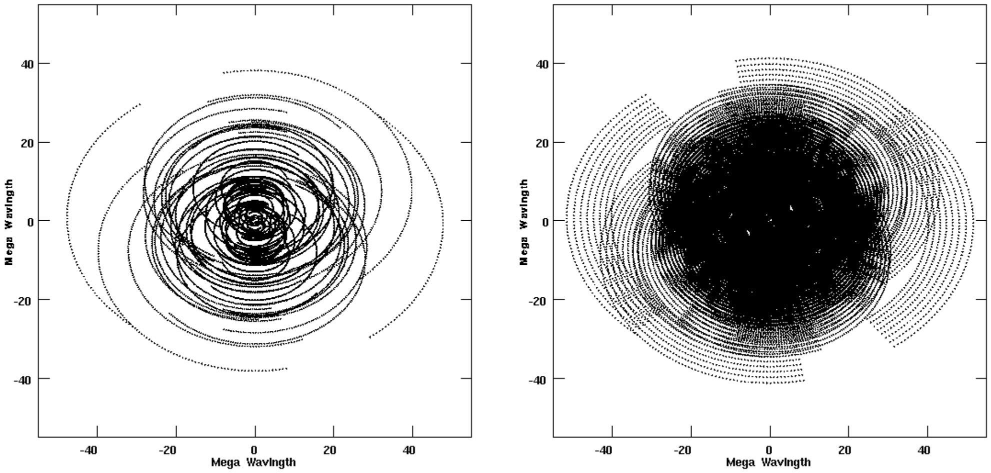


Figure 21–1. *Left (a)*: VLBA (u, v) coverage for a full track at $\delta = 50°$. *Right (b)*: Using MFS observations with eight frequencies spread over 25%.

2. Spectral Variation

One of the reasons that multi-frequency synthesis techniques work is because, for continuum emission mechanisms, physics virtually guarantees that the spectrum at each point on a source is relatively smooth over a significant range of frequencies. For synchrotron emission, a single electron at energy E will radiate over at least an octave in frequency centered at a frequency which depends on E. Furthermore shock acceleration mechanisms give smooth distributions of E. For optically thick thermal emission, the spectrum is given by the Rayleigh-Jeans formula: flux density increases as the square of frequency.

Often the spectral variation of a source is assumed to be a power-law relationship, with spectral index α:

$$I(\nu) = I(\nu_0)(\frac{\nu}{\nu_0})^{\alpha}. \tag{21–1}$$

For general spectral variations, we will use the equivalent spectral index

$$\alpha_{\rm E} = \frac{\nu}{I}\frac{\partial I}{\partial \nu}. \tag{21–2}$$

For sources with power-law variations, this is the same as the normal spectral index.

3. The Linear Spectral Variation Approximation

Our aim here is to determine the effects of spectral variation, which will ultimately allow us to correct for this. While there are a number of approaches to deriving the result of this and the following section, here we will assume that the emission of all the sources in the field varies linearly with frequency. We do *not* assume that all the sources vary in exactly the same fashion. For example, some sources in a field may have a very flat variation with small spectral slopes, whereas others might be steep, with large slopes. Plainly a linear variation will be a good approximation if the frequency range is small enough. Later we will generalize the spectral variation to other than just a linear one.

The spectrum of a source can then be written as

$$\begin{aligned} I(\nu) &= I(\nu_0) + \frac{\partial I}{\partial \nu}(\nu - \nu_0) \\ &= I(\nu_0) + \nu_0 \frac{\partial I}{\partial \nu}\frac{(\nu - \nu_0)}{\nu_0} \\ &= I(\nu_0) + \alpha_{\mathrm{E}} I(\nu_0)\frac{(\nu - \nu_0)}{\nu_0}. \qquad (21\text{–}3) \end{aligned}$$

Here ν_0 is some reference frequency, which will normally be near the center of the frequency range of interest.

We can now consider making an image of a point source, at the phase center, using multi-frequency data for some given (u, v) coverage. If we know, *a priori*, what the spectral variation of the point source is, then we can readily predict the image that we would make from such data. Here we assume a conventional imaging algorithm, where no special regard is given to the frequency at which a particular visibility is measured (apart from using the correct frequency when computing the (u, v) coordinate of the visibility). We first consider two extreme sorts of point sources:

- If the source had unit flux density and a flat spectrum, then the image would be just the normal synthesized beam (or point-spread function) – which we represent by $B_0(\ell, m)$. We can form this image by replacing the visibility data by 1 in the imaging step.

- We can also imagine a very non-physical point source, which has a flux density at ν_0 of zero, but where $\alpha_{\mathrm{E}} I = 1$. That is, we form a multi-frequency synthesis image of a point source whose intensity is $(\nu - \nu_0)/\nu_0$. We represent this response as $B_1(\ell, m)$. We form this by replacing the visibility data with $(\nu - \nu_0)/\nu_0$ in the imaging step. Conway et al. (1990) called this response the "spectral dirty beam".

An arbitrary point source (but one which we still assume to vary only linearly with frequency) will be a weighted sum of these two extreme point source types. So the response of an arbitrary point source will be the weighted sum of the responses to the extreme sources. That is, the dirty image, $I_{\mathrm{D}}(\ell, m)$, will be

$$I_{\mathrm{D}}(\ell, m) = I(\nu_0) B_0(\ell, m) \; + \; \alpha_{\mathrm{E}} I(\nu_0) B_1(\ell, m). \qquad (21\text{–}4)$$

For an arbitrary source (i.e. not just a single point source), the dirty image will become a sum of convolutions of the two different responses:

$$I_{\mathrm{D}}(\ell, m) = I(\ell, m) * B_0(\ell, m) \quad + \quad (\alpha_{\mathrm{E}}(\ell, m) I(\ell, m)) \; * \; B_1(\ell, m). \qquad (21\text{–}5)$$

Here we drop the explicit dependence on reference frequency, ν_0. Equation 21–5 shows that the dirty image formed from multi-frequency synthesis data consists

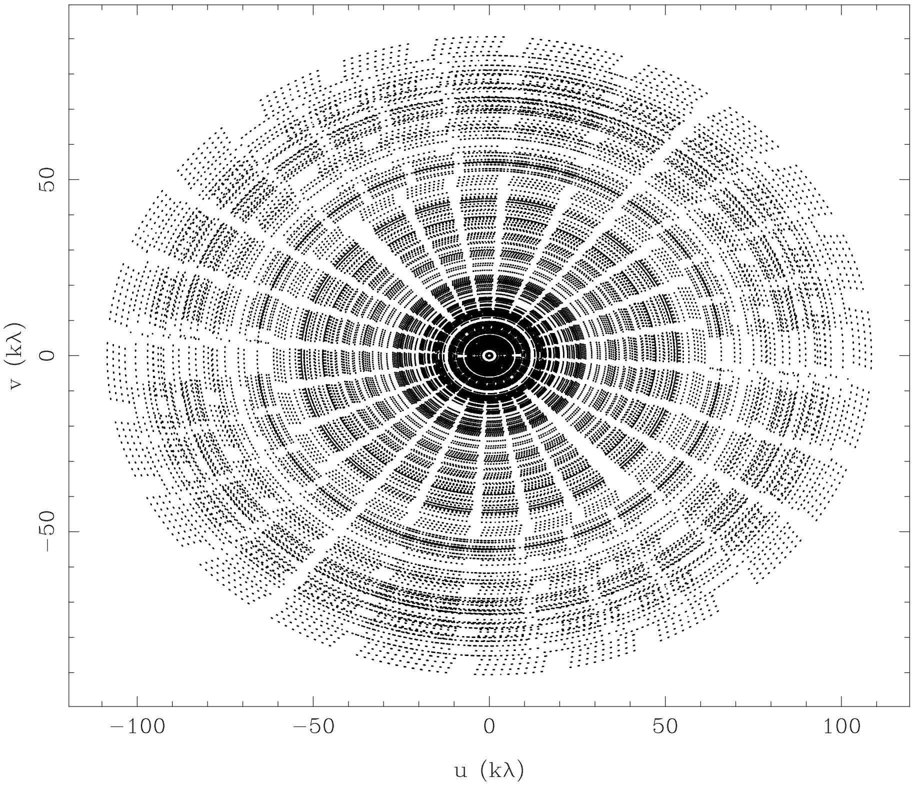


Figure 21–2. The (u, v) coverage for a 14-frequency ATCA observation

of two components – the "normal" response of the emission plus a spectral response. The spectral response component can be thought of as an artifact.

Ideally the spectral response would be zero – which it clearly is if there is no spectral variation ($\alpha_{\rm E} = 0$). As spectral indices are of order 1, the magnitude of $|\alpha_{\rm E} I|$ and $|I|$ are typically similar. So we see that the importance of the spectral response is related to the magnitude of the spectral dirty beam. However, for frequency ranges of about 30%, the response pattern in the spectral dirty beam is typically only 0.01 or 0.02. For example, Fig. 21–2 shows the (u, v) coverage from an Australia Telescope Compact Array (ATCA) multi-frequency observation. Here 14 frequency bands were observed in the frequency range 4.418 to 6.099 GHz (a frequency range of 33%), with each band being 100 MHz wide and consisting of 25 channels (the plot is of only the central channel from each of the 14 bands). Choosing a reference frequency of $\nu_0 = 5.14$ GHz, Fig. 21–3 shows B_0 and B_1 – the dirty beam and the spectral dirty beam – both saturated at ±0.02. Whereas the normal dirty beam has a maximum of 1, the spectral dirty beam has a peak value of about 0.01! The spectral response is two orders of magnitude weaker than the normal response! This is typical – for frequency ranges of 30%, the spectral response is typically about 1% of the main response. Thus for low to modest dynamic range work, the spectral response can be completely ignored.

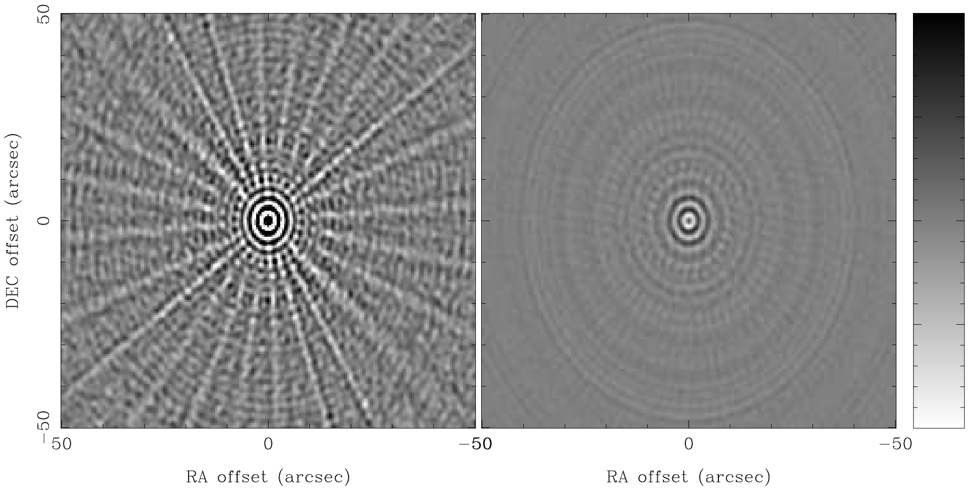


Figure 21–3. *Left (a):* The normal dirty beam, and *Right (b):* The spectral dirty beam, corresponding to the coverage of Figure 1. Both the normal and spectral dirty beams are saturated at the ±0.02 level.

The reason why the magnitude of the spectral dirty beam is relatively weak is easily seen: whereas the normal dirty beam is the image formed by replacing the visibility values with 1, the spectral dirty beam replaces the visibility values with $(\nu-\nu_0)/\nu_0$. For our example, this amounts to replacing the visibilities with values roughly uniformly distributed between -0.14 and 0.19. Apart from being a fraction of 1 to start with, the positive and negative contributions tend to cancel each other out. The choice of the reference frequency is clearly something of some importance – the reference frequency is actually chosen to minimize the response of the spectral dirty beam. In particular, it was chosen so that $B_1(0,0) = 0$. Assuming a linear spectral variation, this results when ν_0 is the (weighted) mean frequency of all the data (the weighting is related to the visibility data weighting).

4. Higher Order Decompositions

While assuming that the spectral variation is a linear one is clearly a good first-order approximation, we will want to be able to generalize this. Although Conway et al. (1990) attack this as a Taylor's series expansion of a power-law function, we will approach it as a general decomposition problem. The previous section assumed the spectral variation was composed of two terms: a constant and a linear variation. The response to these two components could then be determined. In the general case, we can assume that the spectral variation of all sources are decomposable into a sum of some basis functions, $f_i(\nu)$:

$$I(\nu) = \sum_i c_i f_i(\nu). \tag{21–6}$$

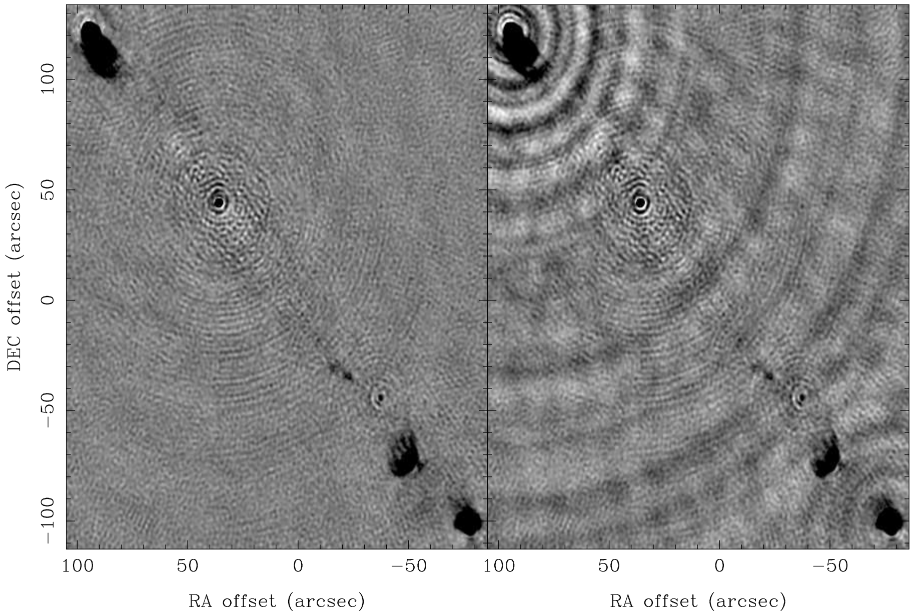


Figure 21–4. The ATCA image of a radio galaxy formed from a multi-frequency synthesis observation. In both cases, the image is saturated at the ±0.15% level.*Left (a)*: Using the multi-frequency synthesis deconvolution algorithm. *Right (b)*: Using a traditional CLEAN algorithm.

We can compute the response image, $B_i(\ell,m)$, of each of the basis functions: $B_i(\ell,m)$ is computed by replacing the visibility data point with $f_i(\nu)$. The multi-frequency synthesis dirty image will be expressible as

$$I_{\rm D}(\ell,m) = \sum_i c_i(\ell,m) * B_i(\ell,m). \tag{21–7}$$

The responses $B_2(\ell,m), B_3(\ell,m), \ldots$ are called higher order spectral dirty beams.

So the "linear approximation" can be seen as a two-term decomposition using $f_0(\nu) = 1$ and $f_1(\nu) = (\nu - \nu_0)/\nu_0$. Conway et al. (1990) have shown that if the spectral variation is really a power law, then for a two-term decomposition, using $f_1(\nu) = \log(\nu/\nu_0)$ actually gives weaker spectral response than the linear approximation.

Are there other good basis functions? Using simple polynomials as basis functions

$$1, \quad \frac{\nu - \nu_0}{\nu_0}, \quad \left(\frac{\nu - \nu_0}{\nu_0}\right)^2, \quad \ldots \tag{21–8}$$

is an obvious choice. Such a choice is equivalent to the Taylor-series approach. Now orthogonality of the basis functions is desirable. This is because, if we assume that frequencies observed are fairly uniformly sampled in some range, then orthogonal basis functions results in spectral dirty beams which have little correlation between each other. Because the simple polynomials are not orthogonal, there is significant correlation between the different spectral dirty beams.

If we assume that the frequencies are fairly uniformly sampled in some range, then the Legendre polynomials (which are an orthogonal polynomial class) can be seen as a good set of basis functions.

The higher order spectral dirty beams have been analyzed in some detail by Conway et al. (1990). They find that the second order artifacts (those resulting from $B_2(\ell, m)$) are at about the 5×10^{-4} level for VLBA-type observations with a 25% frequency range.

5. Calibration and Self-Calibration Issues

So the previous sections show that, for frequency ranges of about 30%, the first order spectral response is typically about 1% of the main response, and the second order response is typically 0.05%. The practical situation is actually even better than this. Assume we know, *a priori*, that the bulk of the emission in the field has a particular spectral variation, $G(\nu)$, then we can calibrate this out of the visibility data. That is, we can generate visibilities compensated for the dominant spectral variation:

$$V'(\nu) = V(\nu)/G(\nu). \tag{21–9}$$

This is like "bandpass correcting" the source for its intrinsic spectral variation.

Unfortunately we will usually not know *a priori* the dominant spectral variation. However, if the data are to be amplitude self-calibrated (and this will usually be so in cases where the MFS image errors at the 1% or so level are of concern), then the spectral variation does not need to be known *a priori.* Assuming that separate amplitude gains are determined at each frequency, then the act of amplitude self-calibration will tend to enforce the dominant spectral index of the self-calibration model. That is, the visibility data are forced to obey the model spectrum. Just as the absolute position and flux density of a source are potentially lost in the self-calibration process, so too can the absolute spectral index. As an example, Fig. 21–4b shows an MFS image (from Sault and Wieringa 1994) where the spectral index of a radio galaxy core was self-calibrated or "flattened" from $\alpha \approx -0.7$ to $\alpha \approx -0.1$ By flattening the core's spectral variation, first and higher order spectral artifacts were eliminated near the core.

In flattening the spectral variation, we assumed that the antenna gains at each frequency are independent, and are solved for independently. Often the gains will not be independent, and it is best to use this to advantage in the self-calibration process. Self-calibration will work better if we constrain the solution process to those solutions which are physically realistic. If we measure simultaneously at several frequencies (or time-multiplex between the frequencies more rapidly than the timescale of typical phase variations), the phase errors for a given antenna at the different frequencies will be related. There will be only two or three phase parameters per antenna: a frequency constant phase offset, resulting from independent electronics, etc; a delay term, $\phi \propto \nu$, resulting from electronics and troposphere; and at low frequencies, an ionospheric term, $\phi \propto \nu^{-1}$. Therefore over each solution interval there are only two or three antenna-based phase parameters that need to be solved for.

The independence or otherwise of the amplitude part of the gain is less clear, as this might be largely instrumental in origin (assuming the atmospheric opacity is negligible). It is conceivable that the correct approach might be to model the antenna-base amplitude gain variations as the product of a frequency-dependent but time-independent "bandpass" function, and a time-dependent but frequency-independent factor.

6. MFS Deconvolution

Often the spectral response can be ignored in MFS images, either because the dynamic requirements are modest, or because the calibration/self-calibration techniques are effective at minimizing it. However this will not be the case for high dynamic range images, where the emission has a variety of spectral indices. In this case, for a first order spectral decomposition, we need to develop an algorithm to "deconvolve"

$$I_{\rm D} = I * B_0 \ + \ (\alpha_{\rm E} I) \ * \ B_1. \tag{21–10}$$

Several algorithms have been developed which are very effective at solving this. We discuss these MFS deconvolution algorithms (in order of increasing complexity) in the following subsections.

All these algorithms deduce a spectral index image, in some form. MFS processing, however, is generally *not* a good way to determine spectral indices. MFS processing deals with comparatively modest frequency spreads, and therefore modest flux differences between the different frequencies. Large frequency spreads (e.g. factors of two or more separation in frequency) are the better approach if the main intent is spectral index information.

6.1. Map and Stack

This is the simplest approach: make individual channel images at each frequency, stack them together to form a cube, and estimate the spectral variation at each pixel from the resulting 'data cube'. Having determined this, the spectral variation is subtracted from the visibility data, which are then Fourier inverted, CLEANed, etc.

Unfortunately, because the spectral dependence is ultimately determined from single frequency images, this algorithm feeds back the large single-frequency reconstruction errors into the data. It is possible to show that some forms of reconstruction errors, such as sinusoidal ripples in CLEAN images (i.e. 'CLEAN ripples'), will not be reduced significantly using this algorithm (Conway 1988). Philosophically we can argue (Conway et al. 1990) that any algorithm which processes the individual frequency data sets piecemeal must be non-optimum since it does not exploit the non-linearity of deconvolution algorithms. However, the algorithm does have the advantage that the variation need not be simple and can be easily found. There may therefore be a role for this algorithm in estimating and removing effects in small bright regions of a source prior to using one of the other algorithms.

6.2. Direct Assault

We could treat the MFS problem as one of explicit model fitting. As we have unknowns $I(\ell, m)$ and $\alpha_{\mathrm{E}}(\ell, m)$ at each pixel, we should vary them to better fit the MFS data set, i.e. to reduce the χ^2 fit to the data. Optionally we can include a a 'regularizing' penalty function P (i.e. such as the sum of the image entropies for the resulting model images at each frequency), and then minimize $\chi^2 - P$. The function P acts to bias the output image to have the properties that *a priori* we know or assume it to have, i.e. positivity, smoothness etc. Although it may be 'using a sledgehammer to crack a nut', with improving computing power it might be the best approach.

6.3. Modified Direct Assault

This is an approach which combines the first two techniques. Here a dirty spectral cube is formed, and a joint deconvolution is performed where the solution is constrained to be spectrally smooth. This approach is very general, and can potentially work in very wideband applications. Indeed, Koom, Hurford and Gary (1997) have used such an approach with the Owens Valley solar interferometer. In their observation, they used three antennas which sampled visibilities at 45 frequencies in the range 1–18 GHz every 10 s. Clearly such a large range of frequencies provide a substantial enhancement in the Fourier coverage. Their approach is the converse of the conventional approach of having large numbers of baselines and small number of instantaneous frequencies. In the imaging step, they formed a spectral cube with the deconvolution algorithm ensuring spectral smoothness. In particular, they used a maximum-entropy-like deconvolution algorithm, where they included a spectral smoothness measure

$$P = -\sum_{\ell,m,\nu} \tau(\ell, m, \nu) \log(\tau(\ell, m, \nu)) \tag{21–11}$$

where

$$\tau(\ell, m, \nu) = 1 + |I(\ell, m, \nu) - I'(\ell, m, \nu)| \tag{21–12}$$

and where $I'(\ell, m, \nu)$ is the intensity interpolated from the two neighboring frequencies at the same spatial position. This technique is likely to play a significant part in future solar interferometer arrays.

6.4. Data Weighting Methods

Cornwell (1984) discussed methods in which spectral effects are reduced or eliminated by appropriate weighting of the data in the uv plane. If a single (u, v) point was sampled at two different frequencies ν_1 and ν_2 then $V(u, v, \nu_1) = I + (\nu_1 - \nu_0)\partial I/\partial\nu$ and $V(u, v, \nu_2) = I + (\nu_2 - \nu_0)\partial I/\partial\nu$. Then $I = a_1 V(u, v, \nu_1) + a_2 V(u, v, \nu_2)$ where the constants a_1 and a_2 depend only on ν_1 and ν_2, thus eliminating spectral effects.

In general, it is unlikely that any two (u, v) points will exactly overlap. However I and $\partial I/\partial\nu$ will be approximately the same for two different visibility points provided they are well within a distance $1/\Delta\theta$ wavelengths in the (u, v) plane where $\Delta\theta$ is angular size of the field over which there is emission from the source (or the primary beam). It is possible to get a large reduction in spectral effects over the field of the source if we incorporate frequency dependent weighting into the 'uv gridding' process. This process grids the irregularly sampled (u, v) points on a square grid of cell size of order $1/\Delta\theta$ prior to Fourier transformation.

Algorithms of this type were briefly considered by Conway (1988). The major problem is that only those grid cells with two or more (u, v) points at different frequency can contribute to the MFS image. For large complex images for which MFS is most needed, $1/\Delta\theta$ becomes small, few cells are double-sampled and so the effective MFS (u, v) coverage is relatively poor. For typical image sizes and MFS observing parameters the effective MFS uv coverage can be comparable to or even worse than, the single frequency (u, v) coverage, thus removing the whole point of MFS! However, since the effective MFS (u, v) coverage is a strong function of image size this algorithm may have a role in producing very high reliability images of relatively small objects.

6.5. Double Deconvolution

This variant of the CLEAN algorithm (Conway et al. 1990) is based on the original justification of CLEAN as a pattern recognition algorithm (see Högbom 1974). The MFS dirty image is formed from the superposition of $I(\ell, m)$ and $\alpha_{\rm E}(\ell, m)I(\ell, m)$ distributions each convolved with its own beam i.e B_0 or B_1. Therefore point components in either the I or $\alpha_{\rm E}I$ distributions have unique signatures in the dirty image. We can try to find such point components in I or $\alpha_{\rm E}I$ by alternately correlating the dirty image with B_0 or B_1 and recording point components (i.e. CLEAN components) at those parts of the convolved image with high brightness.

A significant advantage of "Double Deconvolution" over "Map and Stack" approaches is that the information about frequency sampling is very compact – it is contained within the spectral dirty beam. It can readily handle a very large number of irregularly sampled frequency channels, with each frequency channel potentially being very poorly sampled in the (u, v) plane. This needs to be compared with "Map and Stack" approaches, which require a plane per frequency and implicitly assume that the image at each frequency is modestly good. For example, the ATCA example in Fig. 21–2 used 350 distinct frequency channels, each of which had individually poor (u, v) coverage and sensitivity. Reducing such an observation using "Map and Stack" is not attractive.

In detail, "Double Deconvolution" proceeds in the following way. In the first half of the first cycle we start with the MFS dirty image and attempt to search for the I distribution. Strictly we should correlate the dirty image with B_0, but if the data are uniformly weighted, then $B_0 \star B_0 = B_0$ and $B_0 \star B_1 = B_1$, so we can omit this step. Here '$\star$' represents the correlation operator.[1] Because B_0 has a central peak, and B_1 does not, at this stage the dirty image will be predominantly $I * B_0$. In this half of the algorithm we recognize I components (removing B_0) until the fit no longer improves, giving us an estimate of the I distribution – I_m. After removing the effects of this model from the data the residual image is now

[1] Because the transforms of synthesis imaging beams are usually real-valued, correlation is normally the same as convolution

$$(I - I_m) * B_0 + (\alpha_{\mathrm{E}} I) * B_1 \tag{21–13}$$

In the second half of the first cycle we recognize $\alpha_{\mathrm{E}} I$ components by correlating the residuals with B_1.

$$(I - I_m) * (B_0 \star B_1) + (\alpha_{\mathrm{E}} I) * (B_1 \star B_1) \tag{21–14}$$

Now the I part is convolved with a beam with no central response, whereas $B_1 \star B_1$ will have a central response and so the residuals in this correlated image will be predominantly $\alpha_{\mathrm{E}} I$. We now CLEAN with $B_1 \star B_1$ and form an $\alpha_{\mathrm{E}} I$ model. The effects of this model can now be removed from the data, the I distribution searched for and a new cycle begun. The process of alternately searching for the I and $\alpha_{\mathrm{E}} I$ distributions can be carried on iteratively until the fit no longer improves.

Less qualitatively it can be shown (Conway 1988) that the algorithm always converges to fit the data, and so, in cases in which the MFS (u, v) plane is oversampled, the algorithm will give a unique solution for the two distributions. More generally it can be shown mathematically (Conway 1988) that convergence towards a correct separation is guaranteed as long as we switch half-cycle when the peak residual is larger than the largest sidelobe due to the unwanted distribution. We can attempt to fulfill this condition by monitoring the r.m.s. sidelobe level in a distant region of the image thought, *a priori*, to be free of radio emission. However, we find in practice that the exact point at which we switch half cycles is not very critical because it takes a very large number of CLEAN iterations to incorporate a lot of the emission from the diffuse unwanted distribution.

6.6. Modified Double Deconvolution

Another CLEAN-based algorithm (Sault 1992; Sault and Wieringa 1994) has been developed and used at the ATCA. Although in many respects similar to "Double Deconvolution", it searches for the two beam patterns simultaneously, and is therefore a more natural analog of normal CLEAN. The chief advantage of this algorithm over "Double Deconvolution" is that there is less possibility of confusion between the responses due to the B_0 and B_1 beams, as we fit to both simultaneously.

Using one-dimensional notation for simplicity, for each iteration, normal CLEAN can be viewed as finding a location, j, and flux, a_0, which minimize

$$\epsilon^2 = \sum_i (R(i+j) - a_0 B_0(i))^2 \tag{21–15}$$

(where $R(j)$ is the residual image). Provided uniform weighting is used, this amounts to simply finding the location and intensity of the peak residual. For the modified "Double Deconvolution" algorithm, the optimum location, I and $\alpha_{\mathrm{E}} I$ components are determined for a point source simultaneously. In more detail, a location, j, and coefficients, a_0 and a_1, are found which minimize.

$$\epsilon^2 = \sum_i (R(i+j) - a_0 B_0(i) - a_1 B_1(i))^2. \tag{21–16}$$

Unlike "Double Deconvolution", there is an obvious generalization to cope with higher order spectral effects.

Although the process of finding the optimum point source is not as intuitively pleasing as locating the peak residual, some manipulation shows that the optimum occurs at the location which maximizes

$$R_0(j)^2 A_{11}(0) + R_1(j)^2 A_{00}(0) - 2R_0(j)R_1(j)A_{01}(0) \tag{21–17}$$

where we define the various images

$$\begin{aligned} R_0 &= R \star B_0, & R_1 &= R \star B_1, \\ A_{00} &= B_0 \star B_0, & A_{11} &= B_1 \star B_1, \\ A_{10} &= B_1 \star B_0, & A_{01} &= B_0 \star B_1. \end{aligned} \tag{21–18}$$

Similarly, the coefficients a_0 and a_1 are found to be simple functions of the A, R_0 and R_1 images at location j.

Whereas the A images are auto- and cross-correlations of the beams (and hence constant), R_0 and R_1 are images which are correlations of the residuals with the beams. Rather than perform this correlation process each iteration, R_0 and R_1 can be computed once at the start of the CLEAN, and then updated at each iteration, by subtracting off the appropriate weightings of the A images.

This algorithm is a factor of a few slower than conventional CLEAN. As spectral effects will be unimportant during the early stages of CLEANing (particularly if the visibilities have been compensated for the dominant spectral variation), a speed advantage can be achieved by starting with a normal CLEAN, and then switching to the MFS mode.

Figure 21–4 shows an example of this modified "Double Deconvolution" algorithm (from Sault and Wieringa 1994). This figure gives images resulting from the (u, v) coverage and beams in Figures 21–2 and 21–3. The images, of a radio galaxy with core and north-east and south-west hot-spots, were formed from visibility data which have been compensated by amplitude self-calibration so that the core has a flat spectrum. Thus there are no MFS artifacts near the core. The left panel shows the restored image resulting from the modified "Double Deconvolution" algorithm, whereas the right panel uses a conventional CLEAN algorithm. Both images are saturated at the ±0.15% level. Clearly, the conventional CLEAN has been unable to remove sidelobes from the north-east hot-spot, which has more flux and a steeper spectral index than does the south-west hot-spot. The remaining artifacts in the left panel are predominantly caused by non-closing errors, not MFS.

7. Conclusion

Multi-frequency synthesis provides a way of drastically improving (u, v) coverage, and hence image fidelity. The effects of spectral variation are not important for images of modest dynamic ranges. For high dynamic-range work, calibration techniques and deconvolution algorithms can help eliminate spectral artifacts. MFS allows designers of continuum instruments an alternative to simply building more antennas.

We close this chapter with two areas where further research into multi-frequency synthesis is needed: the fields of mosaicing and of high rotation measure polarimetry.

7.1. MFS and Mosaicing

When mosaicing, or when interested in any field which is appreciable in size compared with the primary beam, the primary beam itself will introduce a spectral variation. This is because the primary beam attenuation at a particular position in a field is a function of frequency. This effect is negligible near the pointing center but increases towards the edge of the primary beam. To estimate the importance of the effect, assume the primary beam response, P, is a Gaussian form:

$$P(\theta,\nu) = \exp\left(-4\log(2)(\frac{\theta}{\theta_0})^2(\frac{\nu}{\nu_0})^2\right). \tag{21–19}$$

Here θ is the angular distance from the pointing center and θ_0 is the primary beam FWHM at the reference frequency. This is equivalent to a spectral index of

$$\begin{aligned} \alpha_{\mathrm{E}} &= \frac{\nu}{P}\frac{\partial P}{\partial \nu} & (21–20)\\ &= -8\log(2)(\frac{\theta}{\theta_0})^2(\frac{\nu}{\nu_0})^2. & (21–21)\end{aligned}$$

At the half-power point ($\theta = \theta_0/2$), and at the reference frequency, this corresponds to an effective spectral index of -1.4.

In mosaicing, the primary beam response is as important a component in the overall point-spread function as the synthesized dirty beam. Similarly, in MFS mosaicing, the derivative of the primary beam response with frequency is as important as the spectral dirty beam. If we have these two spectral responses, then in principle, we can compute the overall response of a MFS mosaic. Although MFS effects seem to be important in some ATCA mosaics, no algorithms have been proposed to date to deal with this type of observation.

7.2. MFS and Faraday Rotation

For linear polarization images, Faraday rotation may need to be considered as a cause of spectral variation. If we assume that the total polarized intensity obeys a power-law relationship (with spectral index α), for a rotation measure R_m, speed of light c, intrinsic polarization angle χ_0 and total linearly polarized intensity p (a real number), then $Q + iU$ will vary as

$$Q(\nu) + iU(\nu) = p(\nu_0)(\frac{\nu}{\nu_0})^\alpha \exp\left(2i(R_{\mathrm{m}}\frac{c^2}{\nu^2} + \chi_0)\right). \tag{21–22}$$

In this case, the equivalent spectral index is conveniently expressed as a complex quantity

$$\begin{aligned} \alpha_{\mathrm{E}} &= \frac{\nu}{Q+iU}\frac{\partial(Q+iU)}{\partial \nu} & (21–23)\\ &= \alpha - 4i\frac{c^2}{\nu^2}R_{\mathrm{m}}. & (21–24)\end{aligned}$$

The magnitude of this effect in the individual Q and U images will be related to $|\alpha_{\mathrm{E}}|$. As an example, for a rotation measure of 30 rad m^{-2} (a typical value for the Galactic component of the rotation measure) and frequency of 1.4 GHz, the imaginary part of α_{E} is $-5.5i$. Clearly this effect can become significant at low frequencies or high rotation measures, even for the comparatively low dynamic ranges typical of Q and U images. For the very high rotation measures possible in radio galaxies, the effects can be tremendous.

References

Conway, J.E., 1988, Ph.D. thesis, University of Manchester.

Conway, J.E., Cornwell, T.J., & Wilkinson, P.N. 1990, *MNRAS*, 246, 490–509.

Cornwell, T.J. 1984, VLB Array Memo No. 324, NRAO.

Högbom, J.A. 1974, *A&AS*, 15, 417–426.

Komm, R.W., Hurford, G.J., & Gary, D.E. 1997, *A&AS*, 122, 181-192.

McCready, L.L., Pawsey, J.L., & Payne-Scott, R. 1947, *Proc. Roy. Soc.*, A190, 357–375.

Sault, R.J., 1992, ATNF Technical Document Series 39.3019, ATNF.

Sault, R.J., & Wieringa, M.H. 1994, *A&AS*, 108, 585–594.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

22. Very Long Baseline Interferometry

R.C. Walker
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. Very Long Baseline Interferometry is the branch of radio interferometry that uses antennas that have no direct link for data and/or clocks. The baseline lengths are limited only by the size of the Earth, and not even that for space VLBI. The resolution is very high — of order a milliarcsecond. This allows imaging of galactic objects with sizes of about an astronomical unit and of extragalactic objects, even at high redshift, with sizes of about a parsec. The wide separation of antennas makes stabilization of phases difficult because of the need for a high accuracy model and because of the very different atmospheres over each antenna. But great progress has been made in the last few years with the advent of the VLBA. For many VLBI observations now, the data reduction techniques and sensitivity are similar to those of connected interferometers. In this chapter, the process of making a VLBI observation is described with emphasis on those areas that differ from connected element interferometer practice. There is special emphasis on techniques for dealing with phase.

1. Introduction

Very Long Baseline Interferometry (VLBI) has traditionally been distinguished from other forms of radio interferometry by the absence of any direct, real-time link between the stations for either local oscillators or for the received signals. Highly accurate atomic frequency standards are used to maintain coherence of local oscillators and wide bandwidth tape recorders are used to record the sampled data. Without the requirement for a direct link, the interferometer elements can be anywhere, allowing the use of baselines nearly as long as the diameter of the Earth, or longer with the use of satellites. The resolution of such an instrument is typically in the neighborhood of a milliarcsecond (mas) — the size of a dime in Los Angeles as seen from Washington D.C. This gives VLBI arrays the highest resolution of any telescopes used in astronomy. Even for sources at the edge of the observable universe, it allows images to be made of regions small enough to change structure on human time scales. Thus a VLBI imaging instrument can be thought of as a movie camera more so than most other astronomical instruments. The high resolution also allows astrometry and geodesy to be done with greater accuracy than with any other technique. The price paid for the use of such long baselines is relatively unstable phases. Also, target sources must have brightness temperatures of a million degrees or more to be detected, so observations of thermal sources are generally not possible.

For a good sampling of the kinds of science done with VLBI, see the proceedings of IAU Colloquium 164 (Zensus, Taylor, & Wrobel 1998). Good references for VLBI technique include the fundamental reference book by Thompson, Moran, and Swenson (1986: hereafter TMS) and the proceedings of two VLBI summer schools, one in Bologna in 1988 (Felli & Spencer 1989) and the other in Socorro in 1993 (Zensus, Diamond, & Napier 1995).

2. VLBI Science

One of the major areas of VLBI research is studies of jets and associated phenomena around compact objects such as those in active galactic nucleii (AGN) and collapsed stars. Collapsed objects provide the necessary conditions for the generation of extremely hot gases, often concentrated in jets with velocities approaching the speed of light. One of the best known results of early VLBI is the observation of apparent faster-than-light motion in such jets — thought to be a projection effect with a relativistic jet moving near the line-of-sight.

Another major area is studies of molecular clouds, star forming regions, old stars, and accretion disks utilizing natural maser emission from a variety of molecules. Such sources can be extremely bright, not because they are hot — typical temperatures are a few hundred degrees — but because inverted quantum level populations are allowing amplification of incident radiation rather than absorption. Such sources can have equivalent brightness temperatures of 10^{15} K or more. Perhaps one of the finest VLBA results so far was the observation of water masers in a beautiful keplerian disk in the nucleus of NGC 4258 (Miyoshi et al. 1995) — a result that provides some of the strongest evidence yet found by any method that black holes exist in the centers of active galaxies.

By being very careful about the geometric model and making accurate measurements of the relative arrival time of wavefronts at different antennas, the VLBI technique can be used to measure absolute source positions to tenths of a milliarcsecond and station locations to a few millimeters (Sovers, Fanselow, & Jacobs 1998). The source catalogs so generated form the basis for the most accurate reference frames used in astronomy (Ma et al. 1998). The station measurements are used to study plate tectonics, Earth rotation, polar motion and other phenomena of interest to geophysics. While GPS is now achieving accuracies close to VLBI and is displacing VLBI for geodesy for cost reasons, VLBI is still the only geodetic technique in use that has something approximating an inertial reference frame — the distant quasars. Therefore only VLBI can track the long term orientation of the spin axis and rate of rotation of the Earth.

3. VLBI Systems

The world of VLBI is far from monolithic. There are nearly 60 different antennas in the stations catalog used by the VLBI scheduling program SCHED, and that catalog is not claimed to be complete. There are several tape recording systems in use and at least one, and sometimes more, correlators per tape system. Some of the most common systems are sufficiently compatible to allow joint use and there are facilities for translating tapes between some of the types when necessary.

There are some important groupings of antennas involved in VLBI. The most obvious is the VLBA, which is a fully integrated instrument with 10 identical antennas, central operations, maintenance, and management, and a 20 station correlator in Socorro, NM (Napier et al. 1994). It replaced the now defunct U.S. VLBI Network. The European VLBI Network (EVN) is the most active group of separate observatories involved in astronomical VLBI. Much of the EVN data are processed at the MPIfR in Bonn, but that will change soon when the 16 station correlator under construction at the Joint Institute for VLBI in Europe (JIVE) is completed. The geodetic community has a number of antennas devoted to VLBI that are observing nearly as much as the VLBA. Their data are processed at the Washington Area Correlator, Haystack, and Bonn. The Australia Telesope has a VLBI component and correlator with the distinction that it can see the southern sky. A number of antennas in the eastern hemisphere are part of the fledgling Asia Pacific Telescope (APT). Russia is building the Quasar Network. NASA's Deep Space Network has an active VLBI program related to spacecraft navigation. Observatories involved in millimeter VLBI have formed the Coordinated MM VLBI Array (CMVA). Many of the above groups, along with unaffiliated antennas, can observe together in "global arrays". The most notable examples are the joint VLBA and EVN global sessions of 2 to 3 weeks that occur 4 times per year.

Early VLBI systems were based on computer or video tape recorders and could utilize bandwidths from a few hundred kHz to 2 MHz. The recording systems in common use today have bit rates in excess of 100 Mbps. The current VLBA and Mark IV systems used by the VLBA, the EVN, the CMVA, and most geodesy stations, are evolutionary enhancements of the older Mark III system. All are based on the same 1 inch instrumentation tape recorders. Recordings made in appropriate modes on one system can be played at the correlators for the others. The VLBA normally records at 128 Mbps, although the tape recorders can handle 256 Mbps each and there are two at each site that can be used in parallel. The Mark IV recorders have a faster maximum bit rate per track and will eventually have two heads for a maximum bit rate of 1 Gbps.

Three other recording systems have appeared. Two are the similar K4 and VSOP systems from Japan that are based on instrumentation cassette recorders. The other is the S2 system from Canada that is based on a stack of 8 commercial grade VHS recorders and is used at a variety of observatories around the world, including on the VLBI segment of the Australia Telescope. The VSOP and S2 systems were developed partly in support of the space VLBI projects that will be discussed by Ulvestad in Lecture 26.

4. VLBI Phases

VLBI observations are not fundamentally different from observations on connected interferometers like the VLA. Most of the same signal handling and data processing steps occur for both, although the mix of steps done automatically by the on-line systems and steps that must be done by the observer is somewhat different. Operationally, VLBI is distinguished by the use of tapes and delayed correlation. But by the time the observer gets the data, that should not be important.

For the observer, one of the most important differences between VLBI and typical connected interferometry is that the phases in uncalibrated VLBI data are less well behaved. This is because the lines of sight from each antenna pass through totally uncorrelated atmospheres (ionosphere plus troposphere). Also, many subtle geophysical and Earth orientation effects are too small to cause problems for the connected interferometers, but are large compared to the resolution of VLBI. With the typical geometric models in use at correlators built prior to the VLBA, the residual phase, delay, and rate variations were usually so high that phase referencing was not possible, or at least required heroic efforts. The common wisdom was that, to detect a source with VLBI, it must be detected within a coherence time, which is usually a few minutes. That detection must involve a solution for delays and rates, in addition to phases — a process called fringe fitting.

Many of the difficulties with the phases can be overcome by using accurate geometric models and by careful observing. This can greatly simplify the use of traditional VLBI calibration methods and makes possible the use of calibration techniques for many VLBI observations that are more typical of connected interferometry. In particular, phase referencing becomes possible, which allows derivation of absolute positions and allows detection of sources so weak that they can only be seen in images based on hours of integration.

In the rest of this section, the special procedures developed to deal with VLBI phases will be discussed. Then, in the following section, all the steps involved in a VLBI observation will be described with emphasis on those areas that are different for VLBI.

4.1. The Model

A correlator must cross multiply signals, from different antennas, that correspond to the same arriving wavefront. But the antennas are at different distances from the source so the wavefront arrives at different times (the delay). Also they are moving at different speeds along the direction to the source, causing different doppler shifts (the fringe rate). An estimate of these time and rate offsets is removed in the correlator hardware. The estimate is derived using a geometrical model. Table 22–1 gives an overview of most of terms that affect VLBI delays at the level of a cm or more. Most are included in the model used by the VLBA correlator. The observer doesn't really need to know these details, but might find it interesting to see some of the subtle effects that are involved.

The VLBA correlator uses a delay model and accurate source and station positions from the geodetic community and clock offsets and rates based on long term GPS measurements. Model errors, after removal of a clock offset and rate, should normally be dominated by the atmosphere. This significant increase in model quality over previous practice has stabilized the raw phases to the extent that delays, rates, and often phases can be calibrated using calibration sources without having to improve the model in postprocessing. The traditional need to fringe fit all sources and to obtain detections within a coherence time of a few seconds or minutes still applies only at extreme frequencies where the atmospheric effects are too strong to allow referencing.

Other than atmosphere and clock offsets, it is possible for the model to be good to a few centimeters at worst for most stations. This should introduce only a very small number of turns of residual phase over many hours. Clock offsets are the largest term in the residuals from the correlator, with typical values of around 50 ns and much higher values possible. These are easy to correct to the nanosecond level because they don't depend on source position and, other than a possible linear drift, are fairly stable (although short term variations are of concern in geodetic/astrometric observations). The atmosphere is another matter. Because of the $\sec(Z)$ scaling, it can vary a lot between

Table 22–1. Terms of a VLBI Geometric Model [a]

Item	Approx max Magnitude [b]	Time scale
Zero order geometry.	6000 km	1 day
Nutation	~ 20"	< 18.6 yr
Precession	~ 0.5 arcmin/yr	years
Annual aberration	20"	1 year
Retarded baseline	20 m	1 day
Gravitational delay	4 mas @ 90° from sun	1 year
Tectonic motion	10 cm/yr	years
Solid Earth Tide	50 cm	12 hr
Pole Tide	2 cm	~1 yr
Ocean Loading	2 cm	12 hr
Atmospheric Loading	2 cm	weeks
Post-glacial Rebound	several mm/yr	years
Polar motion	0.5"	~ 1.2 years
UT1 (Earth rotation)	Random at several mas	Various
Ionosphere	~ 2 m at 2 GHz	seconds to years
Dry Troposphere	2.3 m at zenith	hours to days
Wet Troposphere	0 – 30 cm at zenith	seconds to seasonal
Antenna structure	<10 m. 1cm thermal	—
Parallactic angle	0.5 turn	hours
Station clocks	few microsec	hours
Source structure	5 cm	years

[a] Adapted from Sovers, Fanselow, & Jacobs 1998

[b] For an 8000 km baseline, 1 mas ↔ 3.9 cm. ↔ 130ps

sources, especially at low elevations. And both the ionosphere and the wet component of the troposphere are highly variable in time. See TMS for an extensive discussion of these contributions and see also Lecture 28. Just to give an idea of the magnitude of the effects, the wet troposphere contributes up to about 1 ns of excess delay at the zenith, independent of frequency. The ionosphere can contribute as much as 60 ns or more at 1.4 GHz and scales with ν^{-2}. It can vary by 5% on short time scales and by an order of magnitude between day and night. The VLBA correlator only uses a seasonal and latitude dependent troposphere model and no ionosphere model. Currently there are also some problems at the 20 mas level or so, soon to be fixed, due to a poor nutation model (IAU1980) and out-of-date station coordinates.

4.2. Fringe Fitting

When there is a model error, the measured cross correlation will peak at a value of delay that is offset from the expected value. The interferometer phase, ignoring instrumental offsets, is frequency times delay, so the delay offset will show up as a phase slope as a function of frequency. For example, a delay error of 125 ns will cause a slope of a full turn across a typical VLBA baseband bandwidth of 8 MHz (Usually 4 or 8 baseband channels are used. See Section 5.2. for a definition of a baseband channel.). If the delay error changes with time, so will the phase and there will be a slope of phase with time, called a "fringe rate". Such phase slopes can limit the sensitivity of an interferometer if they limit the bandwidth and time of the interval of data that can be used for detection of a source. Smaller slopes can limit the dynamic range of an image if allowed to degrade the amplitudes when data are averaged. This is because such degradations cannot be described by a product of antenna gains that applies to all baselines — they do not close and cannot be removed by standard self-calibration. For connected interferometers, such phase slopes are typically so small that they are not a concern. For VLBI, mainly because of clock and atmosphere model errors, they cannot be ignored.

The procedure used to find and correct phase slopes is called "fringe fitting". A fringe fit is nothing more than a self-calibration (Lecture 10) that includes not only amplitude and phase gains, but also the derivatives of the phase gains. Normally only the first derivatives are considered, although the second derivative in time is determined for some space VLBI applications. Most fringe fitting programs first transform the data, using FFTs, from amplitude and phase as a function of frequency and time to amplitude and phase as a function of delay and fringe rate. In a multi-baseline fit, the phases from various combinations of baselines connecting a reference antenna to any other can be stacked before the transform to gain sensitivity. Then the peak amplitude point in this space is found. That provides a starting guess for a least squares solution. The fit can be on a per-baseline basis or, like self-calibration, can be global in the sense that all baselines to a station contribute to finding one set of phase, delay, and rate values for the station. Details of fringe fitting are discussed in considerably more depth in Cotton (1995) and references therein.

Weak Sources: Fringe fitting on weak sources can be tricky. The FFT step must find the right peak, or the results will be meaningless. But, with open delay and rate windows, there are a large number of possible values and a detection can only be believed if it is many sigma. Both mm VLBI and space VLBI observers face this problem, the first because of the very rapid phase fluctuations at mm wavelengths and the second because the spacecraft cannot switch back and forth to a calibrator and the position of the orbiter is not as well known as that of a ground station. Both groups have developed sophisticated tools to deal with the problem (see Lecture 26; Rogers, Doelman, & Moran 1995; Lonsdale & Doelman 1998). It is also possible that the the ionosphere will, at times, cause such rapid phase fluctuations at low frequencies that careful fringe fitting is needed. However, with the large primary beams at low frequencies, it might be possible to find a source in the beam that can be used as a calibrator.

Accuracy Requirements: The accuracy required of fringe fit results depends on the goals of an observation. For geodetic experiments, the delays and rates from the fringe fit are the ultimate observable from which the accurate geometric results will be obtained by least squares fit. It is important to get good, high signal-to-noise ratio fit results and to fit all data. Typically after the fit is done, the actual visibility data are no longer used. For imaging experiments, the fringe fit results are used to flatten the phase slopes. For spectral line observations, this allows phase calibration to be transfered between spectral channels and also may allow time averaging, or at least allow the use of longer self-calibration solutions. Often spectral line observations are of strong sources with high SNR, so the best possible calibration is required. For continuum observations, the purpose of the fringe fit is to allow the use of large solution intervals in time and frequency in subsequent self-calibrations and to allow data averaging in both time and frequency to reduce data set sizes. The main requirement is that the residual slopes are sufficiently small that any amplitude loss in averaging is not large enough to degrade the images.

The amplitude reduction suffered when averaging with an incorrect delay is shown in Figure 22–1 for a variety of final average bandwidths. The curves are simply the tops of the $\sin(x)/x$, or sinc, functions that describe averaging with phase slopes. For a project with a peak to off-source rms dynamic range target in the tens of thousands (not uncommon — 10^5 has been reached on the VLBA), such losses should be kept below about 0.3%, the level marked with the dashed line. For such projects, averaging baseband channels together is not desirable so, for typical cases, the averaged bandwidth will be 8 MHz. For that case, the delay accuracy from the fringe fitting should be kept better than about 5 ns. For very weak sources where dynamic range will not be much of an issue, it may be adequate to keep the averaging losses below 5%. But one may wish to average several baseband channels together to gain sensitivity. As can be seen from Figure 22–1, for a total bandwidth of 32 MHz, the 5 ns criterion again applies. For a significant portion of VLBA observations, 5 ns is greater than the full range of delays seen at any antenna. But it may be less than the scatter of delay solutions on weak sources. Therefore, it is probably not just possible, but advisable, to use the fringe fit results from one or more calibrator scans to set the delays for the whole experiment. The main exceptions are likely to be at low frequency, where the ionosphere can contribute large, variable delays, and at any frequency when some *a priori* parameter is poor — such as a source position with an error of more than about 30 mas.

Correlator Averaging Losses: There is an area of concern about averaging losses that is specific to the VLBA. The correlator, for continuum projects, forms spectra of 128 or 256 channels per baseband, then averages, usually to 16. If there is a large delay error — and any clock offsets still apply at this point — there will be some amplitude loss in this average. This loss does not occur in correlators that don't average spectral channels on-line, which includes most that use the XF architecture (Lecture 4). The postprocessing programs know how to correct for this effect if they know the delay with adequate accuracy. For common observing modes, this places less stringent requirements on the fringe fit results than the need to avoid losses in the full baseband average. But, if the correlator averaging produces spectral channels of 1 MHz width or greater, or if there are abnormally large residual delays during correlation, it is possible that the delay accuracy requirement will be driven by the needs of this correction. The only case where this is at all likely to occur is when the 16 MHz baseband channels are used and the on-line averaging is set to the usual default of 16 output channels. It is best to be sure that the correlator delivers output channels of no more than 500 kHz width. Note that the equivalent of the fringe fit for

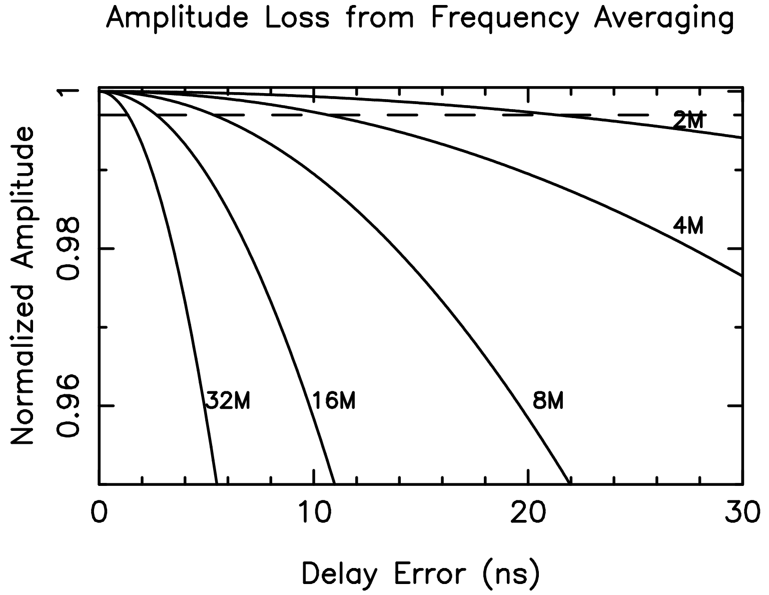


Figure 22–1. The reduction in amplitude for data averaged in frequency the presence of a delay error. The separate lines are for different averaged bandwidths in MHz.

delay could be done with a bandpass calibration, which would find and flatten any phase slopes. But this would not leave a record of the delay offset at the time of correlation and the correction for the correlator averaging loss would not be done. This might actually be acceptable if the *a priori* clocks are very good or if the dynamic range goals of the observation are modest.

Fringe Rates: The need for fringe fitting for the fringe rates should probably be treated separately from the need to fit for delays. At frequencies above where the ionospheric contribution can be large, the fluctuations in delay will be independent of frequency. Their impact just depends on bandwidth which is usually determined by the tape system in VLBI. Just because the phases are much less stable at the highest frequencies does not mean that the delays will be any worse than, say, at 8.4 GHz, so the effort required to determine delays should be more or less independent of frequency. The fringe rates are another story since they scale with frequency. At intermediate frequencies such as 5.0 and 8.4 GHz, with a good model and sources that can be detected in a few minutes or less, it is probably not necessary to fit for fringe rates. Just use self-calibration to solve for phases and use linear interpolation when the calibration is applied. At lower and higher frequencies, removal of rates may be needed. But it is very likely that this can be done using a calibrator near the target source. In all but the most extreme cases, it should not be necessary to fringe fit on a weak source if the observation was scheduled with adequate calibration.

Note that averaging too long in time in the presence of a fringe rate can cause non-closing errors by essentially the same mechanism as averaging in frequency in the presence of a delay error. Figure 22–2 shows the losses as a function of average time and residual fringe rate, assuming a linear phase slope.

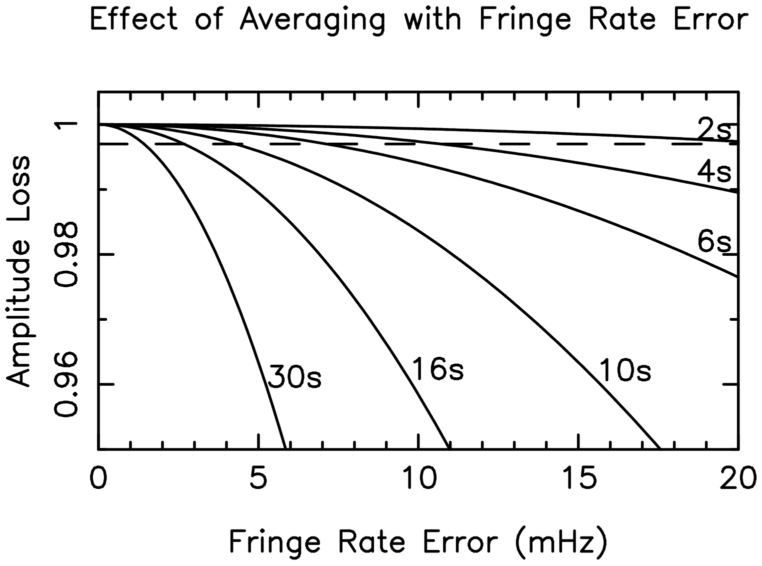


Figure 22–2. The reduction in amplitude for data averaged in the presence of a fringe rate error. The separate lines are for different average times in seconds. Note that loses due to a fringe rate that is known, probably as a result of a fringe fit, will be corrected by the AIPS routines that apply calibration.

A rule of thumb would be to keep the average time below 10 seconds if the residual fringe rates can get as high as 4 mHz, but actual reasonable values will depend on details of the observation. The calibration routines do know how to correct for the amplitude loss caused by a high fringe rate as long as that fringe rate has been determined with a fringe fit.

4.3. Self-calibration

Self-calibration was covered in depth by Cornwell (these proceedings, p. 187) and will only receive cursory attention here. As a quick reminder, self-calibration is the process of using the best available source structure information to improve the antenna phase and, often, amplitude gains. The data set, calibrated with the improved gains, is then used to produce a better image. The process can and should be iterated if the starting model is poor. In fact, such an iterative loop is the standard VLBI imaging method and will be discussed more later.

Self-calibration is based on the assumption that all gain variations can be described as antenna based — they affect all baselines to an antenna equally. Since there are $N(N-1)/2$ baselines with N antennas, and each baseline gives independent information about the source, there is enough information to determine both the gains and the source structure. But as mentioned before, care must be taken to avoid doing something to the data that will cause non-closing errors — amplitude or phase offsets that are not due to source structure and cannot be described by one gain per antenna.

VLBI imaging is totally dependent on self-calibration. First of all, fringe fitting, which is required to make final model adjustments to the clocks, if nothing else, is just a variant of self-calibration. Dynamic ranges beyond about 100, measured as the ratio of image peak to off source RMS, are extremely hard to obtain with phase referencing. If the source is strong enough to self-calibrate, then it is essentially always advantageous to do so, at least for the phases. With self-calibration, dynamic ranges of 1000 are the norm and 10^5 has been reached on the VLBA. Even in cases where phase referencing is required, it is likely that the calibrator will have to be imaged to remove structure effects, since nearly all sources are resolved to VLBI. That imaging step will rely on self-calibration.

4.4. Phase Referencing

In traditional VLBI practice, it has been necessary to detect a source on the baselines to each antenna in an integration time of less than, or equal to, the coherence time. A gain in sensitivity of between 1 and 2 orders of magnitude can be achieved if data from the whole array for the whole time of the observation can be used. This requires calibrating the phases based on something other than the target source. Connected interferometers achieve this goal by observing a nearby calibrator, self-calibrating on that calibrator (that's not usually what it's called, but really that is what it is), and then transferring those phase corrections to the target source. For this to work, it is necessary that the model errors change slowly across the sky so that they are similar on calibrator and target. It is also necessary that the errors change on time scales long compared to the switching time between calibrator and target scans. With the good model on the VLBA correlator, these conditions are often met and phase referencing is possible, even getting to be common. Phase referencing has been done on the VLBA at frequencies between 1.4 and 43 GHz using calibrators less than about 6 degrees from the target source and with intervals of between 15 seconds and a few minutes between calibrator scans.

Phase referencing is limited by whatever position or time dependent model errors remain after correlation and initial calibration. Errors in geometric parameters, such as source and station positions, will degrade phase referenced phases by up to the total error times the calibrator/target separation in radians (i.e., a factor of 0.1 for a separation of 5.7 degrees). Atmospheric effects scale with the secant of the zenith angle, which, unlike the connected interferometer case, can be very different at different array elements for VLBI. Any errors in the steady ionospheric or tropospheric model induce errors scaled by the difference in $\sec(Z)$ between calibrator and target, a factor that can get rather large at low elevations even with close pairs. Atmospheric terms can also vary rapidly in both space and time which will degrade phase referencing using either too long a switching time or too widely separated sources.

The geometric model used by the correlator, ignoring the atmosphere and clocks, should be good to a few cm at worst, which corresponds to less than one to a few turns of phase across the sky (soon to be fixed nutation and station catalog problems cause the current VLBA model to have somewhat larger errors). Clocks are not much of a problem since they do not change with pointing position and can be well calibrated with phase referencing. Also, short term clock variations are generally smaller than atmospheric variations with modern maser frequency standards. But the atmosphere cannot be ignored. It is the dominant source of errors. To understand when phase referencing can be done and with what parameters, we must understand the atmosphere. Unfortunately, both important components, the ionosphere and wet troposphere, can vary tremendously with location and time, so any estimates of how well phase referencing will work must necessarily be very approximate.

Troposphere: The variations in the wet component of the troposphere are especially troublesome for mm interferometry on any but the very shortest baselines. They will be discussed in depth in Lecture 28 based on VLA and other measurements. What has been learned on the VLA about phase fluctuations can also be used to understand VLBI cases. VLBI baselines are always far beyond the outer scale length of atmospheric turbulence. But the longest baselines on the VLA are also beyond the outer scale length and so should not be much different. The characterizations of phase differences as a function of baseline length can be used to understand the differences between calibrator and target lines of sight. Those lines of sight diverge from the antenna, but are actually rather close until beyond the wet troposphere scale height. For a 0.1 radian calibrator/target separation, the lines of sight are only about 100m apart at the zenith at the scale height of about a kilometer. The separation is larger for lower elevation angles. Because of this small separation, except at low elevations, spatial effects are unlikely to be much of a contributer to the relative phase fluctuations. More important are the time fluctuations caused by the turbulence pattern blowing over each antenna. VLA data (Carilli & Holdaway 1997) beautifully show how the fluctuations increase with baseline length as expected. But, if calibration is done on some time interval, the phase fluctuations increase with baseline in the same way until the baseline length is about equal to the distance the wind blows between calibrations. Then the rms phase fluctuations are constant for all longer baselines. There is no reason to believe that the fluctuations would not remain the same, with calibration, all the way out to VLBI baselines. Carilli and Holdaway find that calibration is good to about 5 degrees of phase with a 20 second calibrator cycle and to about 20 degrees of phase with a 300 second cycle time for observations at 43 GHz.

The above discussion would suggest that fluctuations due to the troposphere can be kept small by using a fast calibration cycle. But there is one complication that is far more serious for VLBI than for the VLA. That is that two VLBI antennas will see a source at different elevations. Any error in the total zenith atmospheric delay, not just the fluctuations, and in the mapping function that describes how the delay increases with zenith angle, will show up in the baseline phases. At low elevations, this effect will likely dominate the errors in phase referencing. Note that the same argument applies to both ionosphere and troposphere. Such an error is systematic so it does not beat down with integration.

A future research topic that might help VLBI phase referencing relates to the use of water vapor radiometers to measure the water vapor along the line of sight. Those measurements can be used to correct the interferometer phases. This is currently an area of active development on the mm interferometers and the next generation 22 GHz receivers for the VLA are being built so that they can be used as radiometers.

Ionosphere: The ionosphere is the other significant atmospheric contributor. It can vary by an order of magnitude between night and day. It can also vary with

Table 22–2. Maximum Likely Ionospheric Contributions to Delay and Rate. [a]

Frequency GHz	Max Delay ns	Min Delay ns	Max Rate mHz	Min Rate mHz
0.327	1100	110	12	1.2
0.610	320	32	6.5	0.6
1.4	60	6.0	2.8	0.3
2.3	23	2.3	1.7	0.2
5.0	5.0	0.5	0.8	0.1
8.4	1.7	0.2	0.5	0.05
15	0.5	0.05	0.3	0.03
22	0.2	0.02	0.2	0.02
43	0.1	0.01	0.1	0.01

[a] Adapted from Thompson, Moran, and Swenson 1986

the solar cycle. Table 22–2 gives the maximum likely ionospheric contribution to group delay and fringe rate at 60 degrees elevation for the VLBA frequency bands for quiet (night) and active (day) times. The table is adapted from a table in TMS. Note that the delay contribution scales with frequency squared while phase effects scale with frequency. One quick conclusion from the table is that the ionosphere has a very big effect at low frequencies, and it still can contribute several turns of phase even at 43 GHz. Like the troposphere, the ionospheric contribution scales with $\sec(Z)$ above about 10 degrees elevation. Below that, it starts to level off because of the curvature of the Earth and the fact that the main ionospheric effects occur at an altitude of 300-500 km. Because the delay can be so large, if it is not modeled correctly, the $\sec(Z)$ dependence can lead to very large differences between the model errors on calibrator and target which can be a big problem for phase referencing. The ionosphere is subject to various short term variations including traveling waves. It can vary by 5% due to traveling ionospheric disturbances (TIDs) on time scales ranging from 10 minutes on up. Very short term variations also exist.

The VLBA does not use an ionospheric model of any sort. At lower frequencies, this can be by far the dominant source of model errors. It is likely that phase referencing involving switching the pointing positions of the antennas between calibrator and source will not be possible at 327 and 610 MHz. The best hope of any sort of phase referencing at those frequencies may well be to find a calibrator in the beam when observing a target. Given the large primary beams at those frequencies, and the high density of sources, that might be possible if a large enough fraction of sources are compact. At 1.4–2.3 GHz, phase referencing will not be easy except under low ionospheric conditions, but it has been demonstrated to be possible. Lack of an ionospheric model could significantly degrade phase referencing at even higher frequencies, especially at low elevations.

Global ionospheric models are becoming available from the GPS community. An attempt will be made to apply such models to correct VLBA data in the hopes of improving phase referencing. This is still a research topic, but it has significant promise. GPS ionospheric data is also available directly from dual frequency receivers at some sites. If the global models prove to be of insufficient accuracy, perhaps the local data will be helpful. The ionosphere can also be measured by comparing delays, or with care, phases measured at two well separated frequencies. The VLBA is equipped with special optics that allow simultaneous observations at 8.4 and 2.3 GHz. The geodetic community uses such dual frequency observations to remove the effects of the ionosphere from their geodetic and astrometric observations.

An Example: Figure 22–3 shows an example of phases on both calibrator and target before and after phase referencing. Prior to referencing (top panel), both sources showed phase excursions of several turns, which is typical. This phase winding is due to imperfect models. But the model errors are clearly similar on the two sources because the phases track well. The bottom panel shows what happens when the phases of one source are used to calibrate both. The reference source phases go to zero and the target source shows steady deviations from zero of a few tens of degrees. Such deviations represent some small model error, possibly a source position error. These are test data on two strong sources so it is easy to see how well the phasing worked. The scans were 3 minutes long and the calibrator-target separation was 1.8°. The observations were at 15 GHz.

Scheduling Concerns: Just how should phase referencing observations be scheduled? First, a calibrator needs to be found as close as possible to the target source. It should have adequate flux density to provide reasonable phases during the calibration scans — see the chapter on sensitivity for how to determine this. A few hundred mJy of correlated flux density on the longest spacings is desirable. Weaker sources have been used. A major calibrator survey is in progress that will provide a total of about 3000 calibrators with positions good to a few mas or better and with known visibility functions (Peck & Beasley 1998). That should provide a calibrator that is, on average, about 2 to 3 degrees from the target. Separations up to 5 or 6 degrees can work, but expose the observations more strongly to model errors. The time between calibration scans should be between 15 to 30 seconds at 43 GHz and 2 to 3 minutes at 8.4 GHz (Beasley & Conway 1995). At lower frequencies, longer on-target times are sometimes used, but the risk of encountering problems with turn ambiguities and with ionospheric fluctuations rises. Generally, phase referencing will be easier with shorter times between calibrators, but the total on-target integration time will be reduced if too large a fraction of the time is spent calibrating. It is highly advisable to include a few scans on a phase reference check source in the schedule. This is a source at a similar separation from the main calibrator as the target, but that is itself a potential calibrator. It can be used to check the quality of the phase referencing on that particular day.

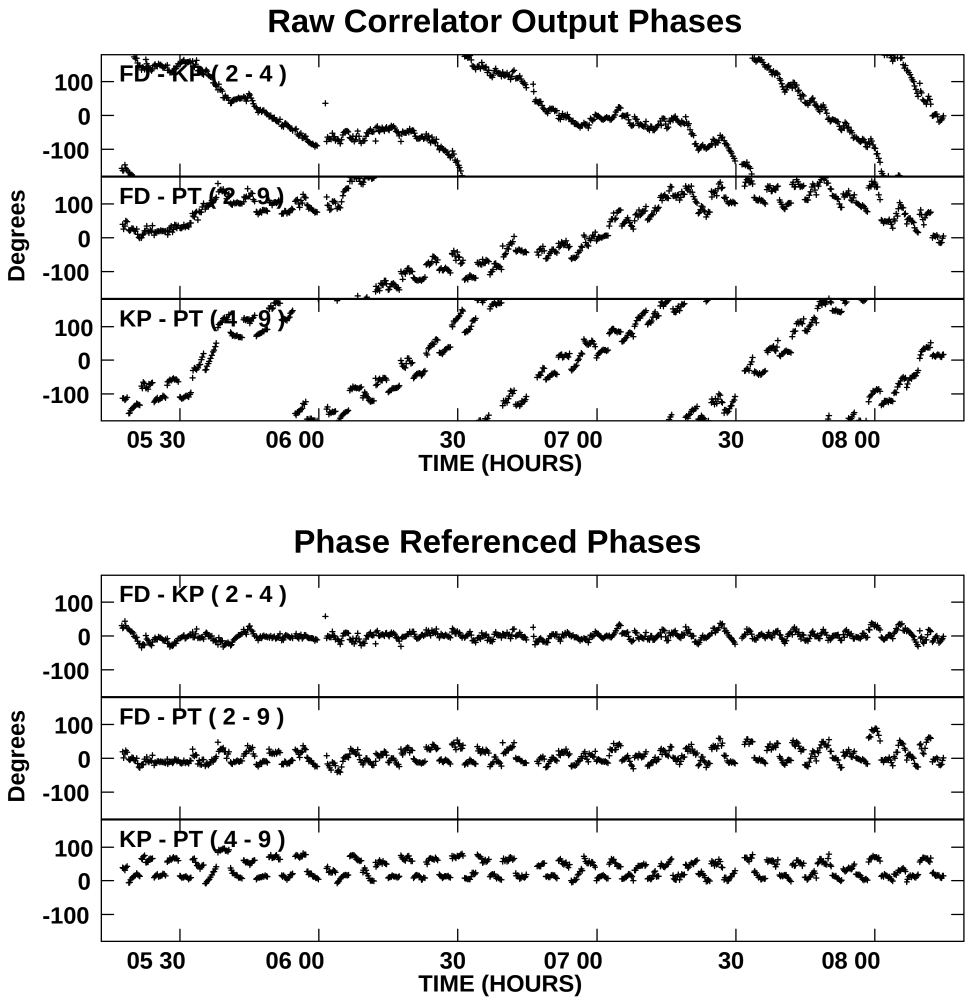


Figure 22–3. An example of phase referencing. The top panel shows raw phases on 3 baselines. The array was cycling between 2 sources, 3C273 and 1222+037, which are about 1.8° apart, with scans of 3 minutes duration. The frequency is 15 GHz. The top panel shows the raw phases with the phase wrapping that indicates that the model is not perfect. But the phases for the two sources track each other well. The bottom plot shows what happens when the phases from both sources are calibrated using the data from one.

5. The Life History of a VLBI Observation

This section is an overview of the steps involved in a VLBI observation from scheduling to imaging. While most of what is presented applies to any VLBI observation, it applies most directly to the case of the VLBA. Special attention is paid to those areas where VLBI practice differs from that typical for connected interferometers. The VLA is used as a concrete example of a connected interferometer.

5.1. Scheduling

Scheduling a VLBI observation involves specifying the individual observation scans and enough information to set up the hardware to the desired configuration. The main VLBI scheduling programs are SKED (for geodesy), PC_SCHED (for Mark III), and SCHED (for nearly everything else including VLBA and Mark IV). Some items to keep in mind while scheduling are:

- Include two or more fringe finder scans on strong sources. Without this, if there are any problems at the correlator, debugging can be very difficult.

- Include at least one or two scans on an "amplitude check source" that is strong and has sufficiently simple structure to be easy to model or image. This source will be used to bring the amplitude calibration of all antennas and basebands to the same scale.

- If observing a weak source, include a nearby calibrator to serve as a fringe rate and delay reference and possibly as a phase reference. If phase referencing, these scans should be at most three minutes or so apart — less at low and high frequencies. For just delay and rate, they can be more widely spaced.

- If phase referencing, include a phase reference check source. This is a calibrator near the main phase calibrator that can be used to check the quality of the phasing on that day.

- If doing spectral line observations (see also Lecture 24), include a bandpass calibrator. It is likely that this will also be the amplitude check source and/or the fringe finder.

- If observing polarization (see also Lecture 25), obtain good parallactic angle coverage on a calibrator at every station to determine the system polarization. The fringe or phase calibrator may serve for this. It is also possible that one observation of an unpolarized source will serve for this.

- If observing polarization, observe something of known polarization to calibrate the polarization position angle.

- If not using automatic tape allocation, which is only available on the VLBA, group scans in units of the length of a tape pass for efficient tape usage. In any case, do not exceed your total tape allocation.

- If using stations with VLBA or Mark IV tape systems and only one tape drive (anything except the VLBA and VLA), allow 10–15 minutes for each tape change, following the guidance provided by the station or network.

- Allow occasional gaps of 2 minutes or more so that the tape systems can do readback checks of recorder health. One every hour or two is adequate.

5.2. Observing

VLBI observations are made based on preset schedules distributed to the antennas ahead of time. The VLBA antennas are usually in contact with the AOC in Socorro, but they do not have to be. For observations involving non-VLBA antennas, there is usually no communication between the antennas during the observations. While it might be possible to make interactive changes to the schedule, it would be difficult and not obviously productive. Therefore, unless the observer is helping with the data aquisition at one of the sites, he/she is usually not involved during the actual observations.

There is very little difference between the antennas used for VLBI and for connected interferometry. In fact, connected element antennas can be, and often are, used for VLBI. Both types focus the radiation into a feed, after which it is converted from circular to linear polarization. Then, a low level, broadband noise signal of known amplitude is injected with some duty cycle. This is used to calibrate the system temperature, which is needed in order to convert correlation coefficients into correlated power. The signals are amplified by low noise amplifiers. Then they are shifted to lower frequencies by being mixed with a reference tone (the local oscillator, or LO), which is derived from the stable frequency standard. After this mix, the signal is called an intermediate frequency, or IF. The IFs are sent to the control building on cables which are subject to bending and thermal variations.

A special feature common in VLBI systems, but not yet in connected interferometers (one is being considered for the MMA), is a second calibration signal that is injected along with the noise calibration signal. It consists of a string of very sharp pulses at, usually, one microsecond intervals that are synchronized to the reference signal from the frequency standard. This has the effect of generating tones of well established phase at one MHz frequency intervals. These tones are subject to all the same instrumental phase variations as the astronomical data. They are detected after the data are digitized and are used to calibrate variations in the delay and phase due to the instrumentation. The VLBA is sufficiently stable that most users don't actually use these tones, but they are valuable for geodesy and as a system integrity check.

The frequency standard is usually a hydrogen maser that is located in the control building. Its output signal has to be sent to the location of the mixers and pulse cal generator, which are near the receivers high on the antenna. The cable involved can change electrical length as it changes temperature or as it is bent when the antenna pointing direction changes. The cable cal system measures the electrical length of this cable and allows corrections to be made. The VLBA, the VLA, and many other systems have cable cals. A difference is that the VLA applies the cable cal on-line while VLBI users must apply it in post-processing. As with the pulse cal, most non-geodetic VLBA users choose to trust the stability of the VLBA and do not apply the cable calibration.

One or more IFs, one for each polarization and perhaps additional ones if multiple frequencies are observed simultaneously, are sent to the VLBI racks in the control building. There each IF is split and sent to several baseband converters ("BBC", which is VLBA terminology; they are called "video converters" in Mark III/IV systems). These units contain tunable synthesizers, also locked to the maser, and filters. They mix the IF down to baseband, where one frequency edge of the signal is at DC. They also limit the bandwidth to any of a number of values, which on the VLBA range in factors of two from 62.5 kHz to 16 MHz. At the VLA, the equivalent step happens in the control building after the signals are sent over the waveguides from the antennas. The VLBA has 8 such BBCs, each able to produce both upper and lower sidebands. Mark III/IV systems have 14 or 16 video converters. One output signal from a BBC is called a "baseband channel" and is the narrowest analog data signal involved in an observation. Usually from 4 to 8 baseband channels are used, often in pairs of right and left circular polarization.

The BBC outputs are sent to the samplers where they are digitized. The digitization is very coarse, either to 2 levels (1 bit) or 4 levels (2 bit). The VLA uses 3 levels. Sampling usually occurs at the Nyquist rate, which is twice the bandwidth. Despite the coarse sampling, the sensitivity loss is not great. The loss relative to an analog system is 0.64, 0.81, and 0.88 for 2, 3, and 4 level samples at the Nyquist rate (TMS). The small number of bits per sample simplifies the electronics. It is also the way to get the greatest sensitivity if the system is limited by the number of bits that can be transmitted, as it is in VLBI with its tape systems. For a given bit rate on the VLBA, 1 bit, 2 level samples give the higher sensitivity, but 2 bit, 4 level samples are only about 2% worse and do it with half the observing bandwidth. For spectral line observations, where the source is of limited bandwidth, this is a great advantage. For continuum observations, narrower bandwidths impose more relaxed requirements on finding delays before averaging and make it easier to avoid RFI. For spectral line observations, there is often excess bandwidth available, even with 2 bit samples. In such cases, faster sampling, called "oversampling", can be used to improve the sensitivity. With factor-of-2 oversampling the sensitivity losses listed above become 0.74, 0.89, and 0.94 (TMS).

Once digitized, the formatter prepares the data for recording. It adds timing information, blocks the data into frames, adds parity bits, and, at least for some VLBA and Mark IV cases, fans out the input bit stream into multiple parallel output bit streams, one for each recording head. A VLBA or Mark IV tape recorder has 32 heads for data and another 4 for system information per head stack and the stack can be moved between passes, allowing recording of 504 tracks total. Mark IV systems will eventually be able to record on two heads. Sometimes the mapping between input data streams and output streams is rotated among several options in what is known as a barrel roll. With this, if a recording head is lost, some of several signals, rather than all of any one, will be lost. By contrast, for non-VLBI observations on the VLA, the digitization happens in the correlator and, of course, tapes are not used for pre-correlation data.

5.3. Correlation

At the correlator, the bit streams are extracted from tape. The gross geometric delay, which can be many milliseconds, is removed using both tape slewing and a buffer. If a bit stream has been fanned out, it is reassembled before correlation. The tapes from all stations in an observation are played back together, maintaining the appropriate time offset to less than one sample. The relative delay can be maintained to better accuracy than the sample interval by introducing a "fractional bit correction", which amounts to applying a phase slope across the frequency band either before (FX correlator) or after (XF correlator) cross correlation. In the VLA, there are delay lines in the correlator that are used to align the signals. The fractional bit problem is handled by having a sampler that can change sample epoch by a fraction of a sample interval.

The antennas are moving toward the target source at different speeds because of the rotation of the Earth (or the motion of the satellite in the space VLBI case). Thus there is a different doppler shift of the source at each antenna. To correlate without very fast phase winding, it is necessary to correct for this doppler shift. At the VLA, this is done by offsetting an LO at the antenna. For VLBI, it is considered safer to limit the complexity of the station equipment, so this correction is done in the correlator by the "fringe rotator". Since this is done after the final filtering, the position of the filter-induced variations in gain across the band shifts with time. This shift has to be accounted for in any bandpass calibration. The "fringe rotator" is effectively a mix with a very low frequency LO. Unlike mixes at higher frequencies, it is not possible to eliminate the unwanted sideband of this mix with filters. Instead, a full complex correlator is used to generate enough information to separate the sidebands and keep only the desired one.

The internal details of various types of correlators are covered in lecture 4 and will be glossed over here. Suffice it to say that the output of the correlator is a time sequence of cross and auto correlation spectra (or lag functions that can be easily converted to spectra). Typical parameters of VLBA correlator output are 16 spectral channels per baseband channel, each averaged for 2 or 4 seconds. Many fewer channels would risk degrading the amplitudes if the *a priori* delays are not very good. Longer averages also risk degrading the data if the *a priori* model is not so good and there are high phase rates. Spectral line observations will use many more channels — between 128 and 1024 per baseband channel in the VLBA case. If very unstable phases are expected or if a wide field of view is needed, average times might be shorter, down to roughly 0.1 second. The combination of average time, spectral points, and number of baselines must not combine to give an output data rate that is too high for the hardware — currently about 500 kbps. By comparison, it is common on the VLA to average a baseband to one spectral point for continuum data and to average in time for 10 to 30 seconds. This level of averaging is possible because of the much smaller geometric and clock uncertainties. Of course, spectral line observations on the VLA can generate lots of channels.

Some correlators, especially those that do geodesy using the Mark III and Mark IV systems (but not the VLBA correlator) also detect the pulse cal tones and fringe fit the data. The exported data are then ready for the geodetic analysis packages. VLBA correlator data must be fringe fitted in AIPS prior to export to the geodetic packages.

5.4. *A Priori* Calibration and Flagging

All of the above steps except scheduling are usually handled by observatory staff. The astronomer is no longer expected to be involved in the actual taking and correlation of the data, at least with the VLBA. After correlation, a distribution tape is written and sent to the astronomer. Also, calibration and log files are made available to the astronomer over the internet. All tasks from this point on are the responsibility of the astronomer, at least in current practice.

The astronomer reads the distribution tape into a postprocessing package — usually AIPS, although in the near future AIPS++ may become a viable option. The processing steps are divided here into 3 stages. The first involves application of calibration and editing information provided by the monitor systems, and perhaps other external sources. The second stage is the use of the data itself to improve the calibration and editing. This may include self-calibration of calibrators and maybe even the target source, but not the final iterative self-calibration and imaging by which final images may be produced. That is the third stage.

There is a considerable amount of monitor data associated with a VLBI observation. It is used in the editing and calibration process. Currently, these data are in ASCII files left on one of a small number of computers, one of which is at the AOC, that are designated for the distribution of files associated with VLBI. The astronomer copies them to his/her home computer by way of the Internet. Special AIPS tasks exist to read the monitor files and fill the relevant AIPS tables. Soon, the data will be provided in tables attached to the main data set at correlation time, so the somewhat tedious step of gathering, preparing, and reading in the flagging and calibration files will no longer be needed.

Flagging: Flag tables allow the user to edit data that the on-line systems knew would be bad. Common reasons include that the antenna was still slewing to source or that an LO synthesizer was not locked to the maser. After application of these flags, typical VLBA data have only a few remaining bad points, so additional editing is fairly easy. Other stations may or may not provide much useful flagging data. A typical global experiment will require considerable interactive editing, although this is improving as more stations provide useful flag information. For VLA observations, the on-line flags are applied at observe time, also providing data sets that are usually quite clean.

Pulse Calibration: The results of the pulse cal tone detection at the sites is another component of the calibration data, at least for VLBA stations. This information may be used to align the phases across the various baseband channels and to remove any delay and phase fluctuations resulting from instrumental effects, such as cable stretching. Tasks exist to load the data and to solve for the phase and delay offsets to apply to the data. The VLBA usually uses two or more tones per baseband which allows phase slopes within each baseband, not just the phase offsets between bands, to be corrected. But this process involves phase ambiguities, which are resolved by the routine that processes the tone data with the assistance of the equivalent of a fringe fit on a reference scan. Then all other scans are assumed to have delays (phase slopes) that are not too different. Cable calibration, discussed earlier, can be applied along with pulse cal corrections. Note that the pulse calibration results are not really needed for routine VLBA imaging observations because of the high stability of the electronics. Other antennas may have more problems that need correction. For the most accurate geodetic and astrometric observations, the pulse cal calibration seems to help even on the VLBA.

System Temperature and Gain: The calibration data also include the system temperatures measured during the observation. These are used, along with gains and gain curves measured at other times, to calibrate the data amplitudes. The basic equation of amplitude calibration is:

$$S_c = b\rho\sqrt{\frac{T_{s1}T_{s2}}{K_1K_2}} \tag{22–1}$$

where S_c is the correlated flux density, ρ is the correlation coefficient delivered by the correlator, T_{si} is the system temperature of the ith antenna, and K_i is the gain (degrees K per Jy) of the ith antenna at the pointing position of the observation (gain curve applied). Typically, system temperatures are provided individually for each baseband channel. The scaling factor b includes both correlator specific scaling factors and corrections for the losses due to the coarse digitization. The term $(b\rho)$ equals the correlation coefficient that would be obtained with a perfect, properly scaled, analog system. For VLBA data read into AIPS with FITLD, corrections have already been applied and the b provided to the calibration tasks should be 1.0.

Gain information for VLBI antennas is generally measured in single dish calibration observations done at times different from the VLBI observations. Gains and gain curves are provided to the user in ASCII files. AIPS tasks exist to read all this information and apply it to the data. Eventually, this information will also be included with the data from the correlator. The provided gains for VLBA antennas are based on large numbers of single dish observations, often spread over years, and can, in the best cases, provide absolute calibration good to a few percent. This is comparable to what is achieved in normal VLA observations using flux density calibrators. Of course, the observer must be alert for antennas with special problems like bad weather and not use them in setting the flux scale.

When the VLA, or Westerbork, is phased up and used as an element of a VLBI array, the local correlation coefficients can be used, along with the flux density of the source, to derive the equivalent of $\sqrt{T_s/K}$ for use in calibration. The required data, and information on how to use it, are provided with the calibration data from an observation.

For non-VLBI observations with the VLA, the system temperatures, or at least something proportional to them, are measured and, usually, applied on-line (the astronomer can choose to block this). This corrects the correlator output amplitudes for many elevation and time dependent effects. Then effectively the gains are measured by observing a source of known flux density. This method could be used for VLBI, but some other instrument would have to be used to make near-simultaneous measurements of the calibrator flux density because essentially all sources that can be seen by VLBI are variable.

Absorption: At the higher frequencies, it is necessary to take into account the absorption by the atmosphere. Some stations provide gain curves that include the effect of absorption, but that is of limited usefulness because the absorption varies, especially at high frequency. Such gain curves are distinguished by a sharp drop at low elevation. The VLBA gains and gain curves do not include the effect of absorption. For the best calibration, the absorption must be dealt with separately. The AIPS tasks involved have this ability based on a number of methods such as provided zenith opacity (from, for example, tipping scans) or using excess system temperature above a provided, or fitted, receiver temperature and spillover. For the best calibration of VLA data, something similar is also needed because the calibrator observations are generally at a different elevation from the target source observations, and the absorption varies with elevation. The AIPS task 'ELINT' can be used to make the appropriate corrections to VLA data.

Polarization: Polarization calibration is covered in Lecture 25 and not much will be said about it here. But even users who have no intention of using their polarization information should be aware of the effect of parallactic angle. As any type of antenna, other than one on an equatorial mount, tracks a source across the sky, the feeds rotate with respect to the source. For observations using circular polarization, this causes a phase shift with time and that phase shift goes in opposite directions for the two hands of polarization. This can cause various problems, including compromising phase referencing. A correction is made for this effect as a standard part of polarization calibration. But, if a polarization calibration is not being made, it is important to at least make this one correction, especially if using phase referencing. Some correlators (but not the VLBA correlator) may do it on-line, but, if not, it is a simple correction that can be computed from the source position, antenna location, and axis type. The AIPS task CLCOR can do the job.

5.5. Data Based Calibration and Editing

After the *a priori* calibration, there may still be stations whose amplitudes are poorly calibrated because of weather, poor calibration data, or other reasons. Also, the phases will not be calibrated, although some alignments may have been handled with the pulse cal data. Further calibration is based on the data itself, either calibrator data or data on the target sources.

AIPS Tables: A word about how AIPS deals with calibration may help some of the discussion from here. AIPS++ will be similar in concept but not in detailed implementation. While calibrating the data, the actual visibility numbers are not modified. All of the calibration and editing information is collected into tables of which there are several types. Some programs use information in one or more tables to derive incremental changes which are put in solution tables. Other programs either interpolate those solutions onto the final calibration tables, or modify the final tables directly. The final calibration tables include the cal (CL) table that includes amplitudes, phases, delays, and rates; the flag (FG) table that has the flagging information; the bandpass (BP) table that specifies the information necessary to flatten the filter responses; the baseline (BL) table which has any baseline specific corrections (not generally needed for VLBA data); and polarization information in the antenna (AN) table. Most programs which read the (u, v) data have options to apply the information in all of these tables. Often there are several versions of some, especially the cal table, that reflect the results at progressive stages of calibration.

Interactive Editing: One of the first data based operations should be to complete the data editing. The monitor-data-based flagging, in a perfect world, would catch everything, and all remaining data would be good. The VLA and VLBA often approach this situation fairly closely. Other systems vary. Conceptually, editing is done by displaying the data in some manner that makes bad points reasonably obvious, identifying those bad points, and adding them to the list of points to be flagged. Operationally, there are many ways to do this. One good thing to keep in mind is that with modern, large correlators, almost all bad points are the result of a station based problem of some sort (including tape playback problems), so you almost always want to flag stations, not just individual baselines. Interactive programs exist that can be used for editing and some programs have the ability to identify and edit points automatically. Also, it is possible, and sometimes more reliable, to use the more general purpose display and flagging programs separately. Editing VLBI data is not terribly different from editing VLA data except that there are usually fewer VLBI antennas, which makes it easier, and the VLBI integration times are shorter, which makes it harder. Also, because even the calibrators are resolved, some of the time-baseline image displays useful for VLA editing are more difficult to use for VLBI.

"Manual Pcal": If real pulse cal data have not been applied, something needs to be done to remove the large phase slopes in frequency that will be there because of clock offsets, and sometimes because of inadequacies in the geometric model. Often it is useful to start with a "manual pcal", which is simply a fringe fit of one scan (sometimes a few scans) whose results are applied to all of the data. This removes any constant clock offsets and aligns the phases and phase slopes across the various baseband channels. It is common to zero the fringe rate (phase slope with time) from this fit, on the assumption that it is likely to be dominated by scan specific effects. This process is sometimes refered to as 'Single Band Delay' calibration. As mentioned before, for a well behaved observation, this may be the only fringe fit that is needed. Note that applying real pulse cal data does the same thing with the added advantage that any time dependent variations in the phases through the system, including relative changes through different electronic paths of each baseband channel, will also be removed.

Example Spectra: Figure 22–4 shows the baseband channel spectra for a VLBA continuum observation. For each baseband channel, there are 16 frequency channels. For each baseline, two baseband channels are plotted side by side. Both the amplitudes (lower) and phases (upper) are shown. The top plots are of raw data. There are phase slopes across the baseband channels and the amplitudes are in units of correlation coefficient. The bottom plots are after amplitude calibration and after a fringe fit on this scan. The amplitudes are now in Jy and the phase slopes and offsets have been removed. Note that the phases haven't actually gotten noiser, but rather the scale is much expanded. The knowledgeable student can probably guess, with a high probability of being right, which source this is from the very high correlated flux density[1].

[1] The source is 3C 84 at 2.2 GHz.

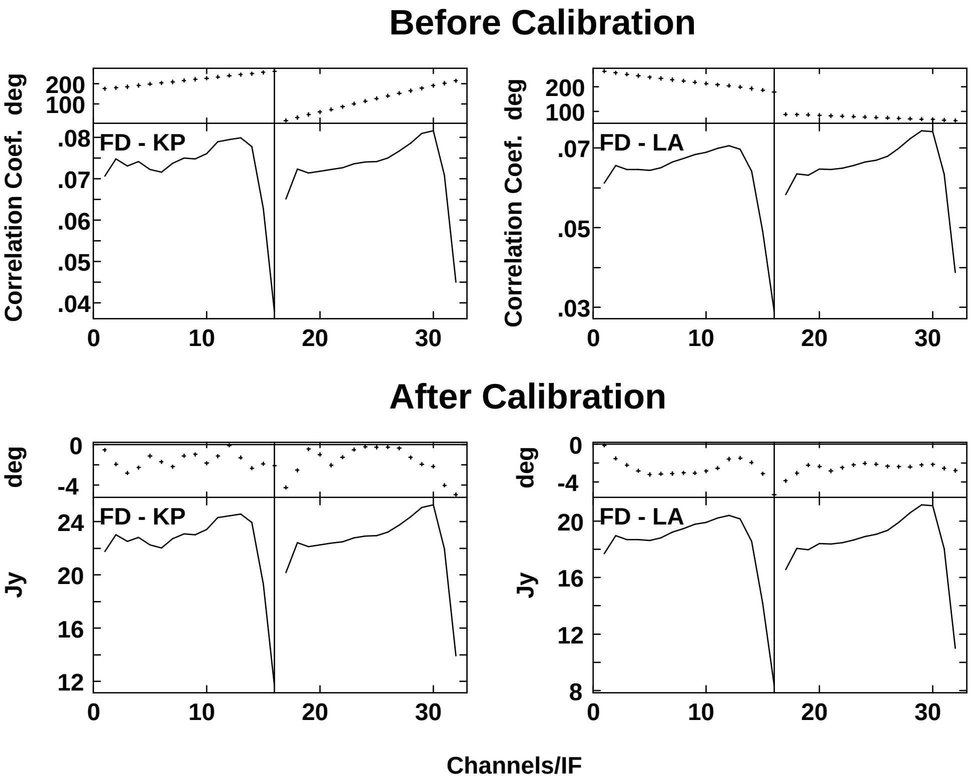


Figure 22–4. A sample of VLBA continuum data before and after amplitude calibration and a fringe fit. For each baseline, both amplitudes (lower) and phases (upper) are shown for two baseband channels. Note the different phase scales before and after calibration — the phases have been flattened.

Fringe Fit: It may prove desirable to fringe fit all the data, or at least all of the calibrator data, to remove variable delays and rates. This can be done before or after the final amplitude calibration, although it must be done after calibration if a source model is used or if one wishes to have the data weights treated properly. When fringe fit solutions are interpolated onto the cal table, the phases are interpolated using the derived fringe rates. That interpolation can go wrong, especially with long solution intervals. When it does, the calibrated data will show a full turn of phase between the times of two solutions, something which can be quite harmful to later processing. If at all possible (source strong enough to tell), such situations should be identified and fixed, which, unfortunately, can be tricky in practice. Or one of the interpolation options that uses phases from integrated rates rather than from the fit results can be used.

Amplitude Adjustments: The *a priori* amplitude calibration is likely to have worked very well for some stations but not very well for others for various reasons, like bad weather or poor available calibration hardware. The “amplitude check source” can be used to fine tune the amplitude calibration for all stations and baseband channels. First an image or model needs to be made — that process is discussed later. Then the model can be used in an amplitude self-calibration step to set the relative amplitude gains of all the antennas and channels. Either in generating the model or image of the calibrator, or after the self-calibration, the amplitudes need to be adjusted so that the average change of gain on the subset of antennas with good *a priori* calibration is unity.

Bandpass Calibration: The data can also be used to do a bandpass calibration. Without such a calibration, there are variations in both amplitude and phase across the individual baseband channels as a result of the imperfect analog filter shapes, as can be seen in Figure 22–4. These can be removed by measuring them on calibration sources using algorithms similar to self-calibration. Bandpass correction is necessary to calibrate spectral line data where you want all channels calibrated individually as well as possible. For continuum observations, in which the spectral channels are averaged together once any phase slopes are removed, a bandpass calibration might improve the accuracy of the closure parameters. But the value will depend on the quality of the filters used in the experiment. For an observation using the VLBA and equivalently good systems, the improvement is probably not great, although it has not yet been quantified. Bandpass calibration can usually be based on a single calibrator scan, although some antennas (such as the VLA) show low-level time variability and it may be necessary to calibrate frequently for the best possible results.

There is a subtlety to bandpass calibration in VLBI that can be important to a decision about whether to use it on continuum data. Because the doppler shifts of the antennas, which are variable, are removed after the final analog filters, the bandpass shape tends to shift around within each baseband. A bandpass shape measured in one scan can only be used to calibrate another scan if it is shifted in frequency by an appropriate amount. Doing so is easy enough, but one is left not knowing how to calibrate one or more of the edge channels, because that region was not measured in the calibrator scan. The current bandpass programs deal with this problem by throwing out the edge channels and for consistency, the same number of channels is thrown out at both ends of the spectrum. Thus a continuum observation will typically loose 2 of its 16 channels per baseband when a bandpass calibration is done. The resulting loss of bandwidth might hurt more than the improvement in the closure characteristics helps. For weak source observations, bandpass calibration is not worthwhile. For very high dynamic range observations on strong sources, it might be.

Spectral Line Amplitude Calibration: For observations of strong maser sources, it is possible to do the relative amplitude calibration using the autocorrelation data. The ratio of the source power to the total system power, including system noise, is what you need to calibrate and is what the autocorrelation provides. The individual autocorrelation spectra, as a function of time, can be fit to a template spectrum to obtain calibration gains. This can be a very effective way to calibrate, especially in the presence of time variable pointing or absorption effects. This method is covered more completely in Lecture 24.

5.6. VLBI Imaging

After the above steps, the data set consists of phases that are reasonably flat with time thanks to good models or fringe fitting. The phases across individual basebands are flat and aligned between basebands. And the amplitudes are in some good approximation of Jy. The data can typically be averaged at this stage to make a file of much more manageable size. For continuum data, usually all spectral channels in a baseband channel are averaged, and for very weak sources, baseband channels are averaged together as well. The time average may be limited by atmospheric fluctuations, but if not, will usually be for something like 10 to 30 seconds. Any longer does not help the data volume all that much and risks various problems such as having a lot of partial records. Of course, spectral line data will usually not be averaged in frequency. The averaged data set is then ready for imaging.

There are two major aspects to imaging - final calibration and actually producing a deconvolved image from the calibrated data. The calibration steps outlined above do not produce calibrated phases for the target source, which are required in order to make a good image. In VLBI, this has traditionally been done almost exclusively through self-calibration, as opposed to the usual connected-element practice of transferring phase from a nearby calibrator. But, as noted earlier, the high quality models now available are changing this. Phase referencing has become reasonable over a wide range of VLBI observations as long as a nearby calibrator can be observed often.

Phase Referencing: The process of transferring phase from a calibrator is the same as for the VLA (Lecture 5). One complication is that it is much more likely that the calibrator will have to be imaged before being used to derive the calibration phases for the target source. However, for observations of a very weak target source with limited dynamic range possibilities, the structure phases on mildly resolved calibrators are not likely to be a big problem and can often be ignored. Another difference from VLA practice is that, especially at extreme frequencies, it is possible that the phase change between calibrator scans will be greater than 180 degrees. This requires utilization of fringe rate information from fringe fits on the calibrator to resolve turn ambiguities.

Figure 9–3 of Lecture 9 on Sensitivity shows an example of a phase referenced image from the VLBA at 8.4 GHz. It is of the radio source in the nucleus of the Seyfert galaxy NGC 5548, a source of about 2 mJy. The image has an off-source rms of 90 μJy.

Phase referencing has the advantage that it can be used on sources much too weak to detect in a coherence time - the maximum time over which self-calibration (including fringe fitting) can be done. It also provides an absolute position, or at least a position relative to the reference source. Such a position can be used for aligning different frequency images or images made at different times. But, the accuracy of the phase transfer will be limited and hence the quality of image that can be made will also be limited. If the target source is strong enough for self-calibration, it will almost always be possible to significantly improve the image quality. It is also possible to make images with self-calibration without the first step of the phase referencing — in fact, this is the normal procedure for strong sources.

Self-Calibration: The process of making an image of a source that is strong enough to to detect in a coherence time, but that has very poorly calibrated initial phases, involves an iterative procedure in which both the source structure and the antenna calibrations are determined. This procedure has sometimes been called "hybrid mapping", but that term is going out of favor. It is now usually just called self-calibration. There is some room for confusion, because that term is also used for one of the steps of the procedure, the one in which the antenna calibrations are determined. The usage in any given instance will have to be determined from context. Self-calibration is covered in more depth in Lecture 10.

The procedure is fairly simple in concept. You start with some sort of initial model or image — a point source works fine, but if something better is available, the process will take less time to converge. That model is used to self-calibrate the data. Don't require very high quality fits at first - you won't get them. The self-calibrated data are then used to make a new image, which hopefully is better than the original. That image is then used for the next round of self-calibration. For the first several iterations (typically 2 to 10), the amplitudes should be held fixed in the self-calibration. The *a priori* amplitude calibration is almost certainly better than what is suggested by the initial poor model. For weak sources and well calibrated data, amplitude self-calibration may not be justified. But for many cases, once progress with phase only self-calibration stops, the amplitude calibration can be added. For sources of modest complexity and with decent data, 10 to 30 iterations might be required to produce a final image and the process is straightforward.

The images from a sample self-calibration sequence are shown in Figure 22–5. This was a simple source with reasonable (u, v) coverage, so the convergence was quick.

Some people have devloped automated imaging procedures which cycle through the self-calibration loop and produce a final image without human guidance. Such procedures have produced something on the order of 90% of the images for some of the larger VLBI surveys. But they don't yet work well for very complicated sources at high dynamic range. Such cases require considerable hands-on guidance. Developing a reliable automatic imaging procedure for VLBI would be a good thesis project for someone interested in algorithms and would be a great boon to VLBI observers.

Self-calibration of large, complicated sources can take time and patience. Large numbers of iterations may be required and every few iterations, one or more parameters of the process needs to be changed or progress stops. The basic problem is probably that, for a source covering a very large number of beam areas, the limited (u, v) coverage of a VLBI array (compared to the VLA for example) just doesn't have enough sampled (u, v) points to specify robustly the source structure and all antenna gains. The self-calibration loop is basically a fitting procedure, and in these circumstances, it keeps getting stuck in local minima and needs to be pushed out of them to resume progress toward the global minimum. Parameters that can be tweaked between iterations to do this include the clean window, the robustness, the taper, the (u, v) range for self-calibration, and the parameters of any clean component editing that is being done. One should calculate the expected noise level, or determine it from something like a stokes V map, and not be satisfied until the off-source RMS is reasonably close. This process is not particularly difficult, it just takes time.

Figure 22–6 is an example of what can be done with persistence using just VLBA data on a low declination source. It shows 3C 120 at 1.6 GHz based on

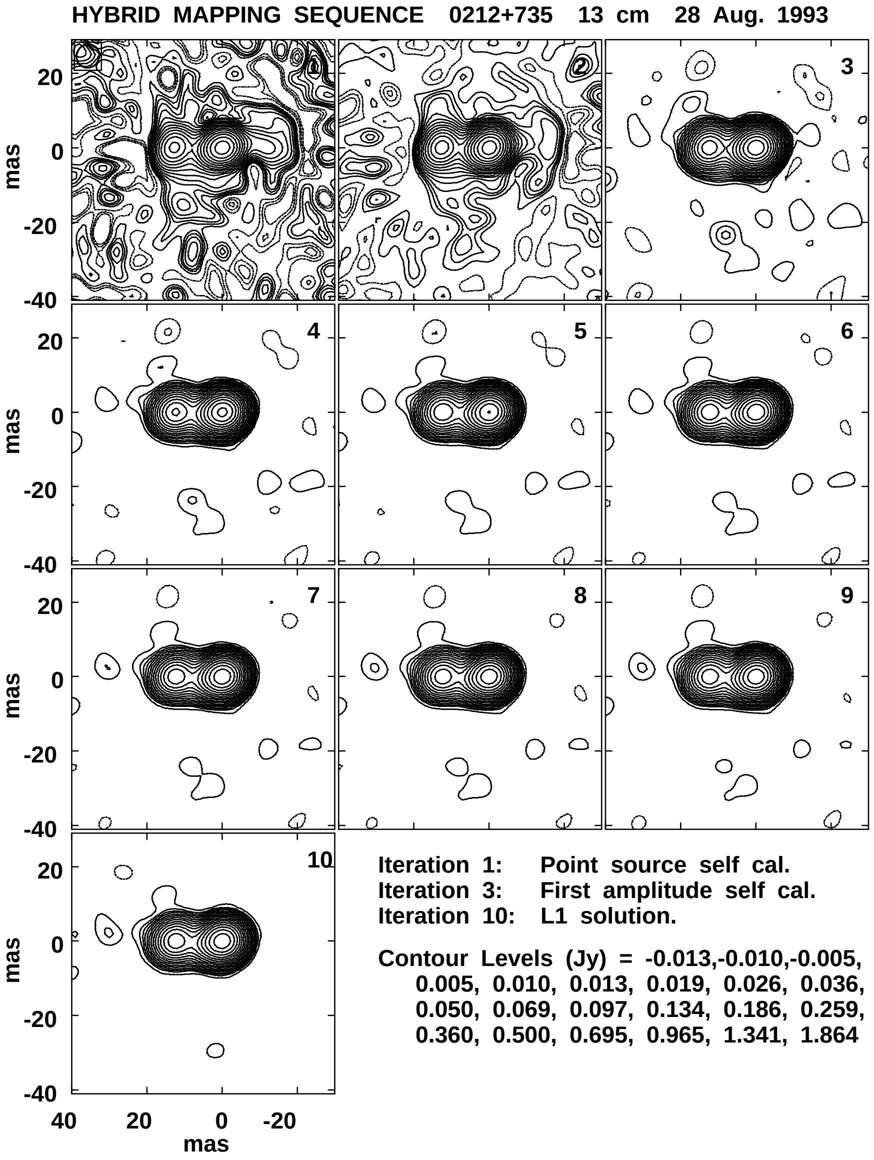


Figure 22–5. The images from each step of a self-calibration sequence. The data are hourly snapshots on a circumpolar source at 2.3 GHz, taken during a geodesy observation. The first two iterations used phase only self-calibration, starting from a point source. The rest used amplitude and phase self-calibration.

data taken in June 1994. The resolution is 7 by 15 mas, extended north-south. The jet is followed out past 0.5 arcseconds.

It is useful, early in the self-calibration process, to explore the beams obtained with a few values of robustness and taper (Briggs 1995). Sidelobes in VLBI tend to be high because of the limited number of baselines. With uniform weighting, the innermost sidelobes tend to be especially large, which can get confused with near-in structural features. With natural weighting, the central

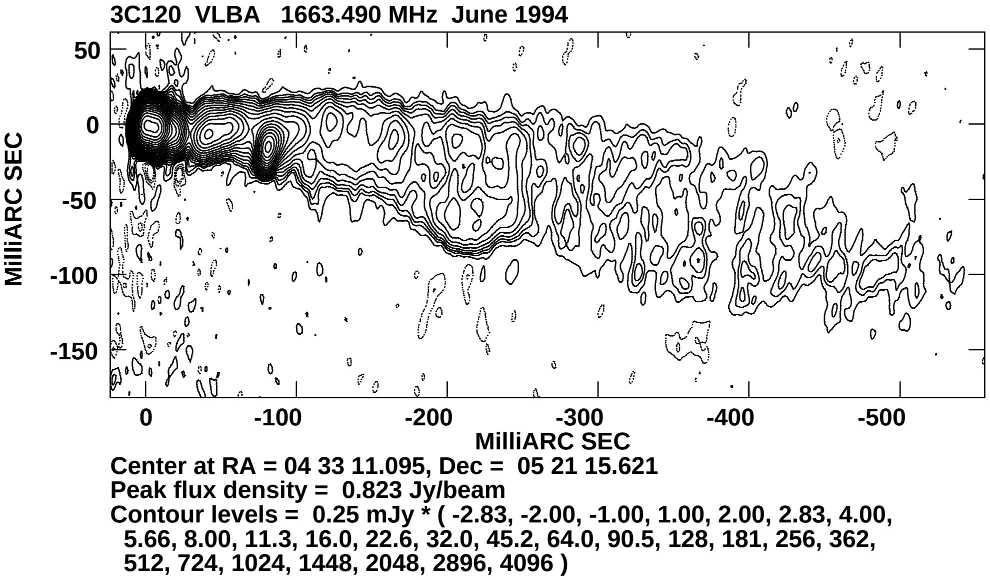


Figure 22–6. A VLBA image of the superluminal radio source, 3C 120, at 1.66 GHz. This demonstrates what can be done with VLBA data on a complicated source.

condensation of baselines, especially with the VLBA, can give a large platform extending well away from the central peak of the beam. This can inhibit proper imaging of large resolved features. It is often possible, with a robustness between −1 and 1, to get rid of the big platform without seriously enhancing the close sidelobes.

One should not use too large a box when using the CLEAN algorithm for the imaging. With limited (u, v) coverage, CLEAN can manage to describe the noise in the data with a modest number of points scattered around a large image and the computed rms will be artificially low. One easy way to tell if this is a problem is to plot a histogram of the values in an image (see Lecture 9). The main peak should be centered on zero and should look Gaussian. If it is a sharp peak with wide wings, you have used too large a CLEAN box. This is mentioned here because it is a common error in VLBI.

Remember that good self-calibration requires that all of the source be described by the model, which is typically a list of CLEAN components. For this to be true, the CLEAN needs to be deep. A shallow CLEAN would leave a significant amount of source flux density out of the model and so it would not fit the data well. That said, in the early iterations, it is typically best to only do a shallow CLEAN, or at least only use the first few points after a merging of components at each position, since any lower level points are likely to be spurious. It is probably better to have an incomplete model containing only real points than a more complete one containing a lot of spurious points.

In AIPS, the self-calibration procedure can be done using a RUN file that strings together separate CLEAN component editing, self-calibration, imaging, display and statistic gathering tasks. An alternative is to use the task, SCMAP, which integrates most of these steps and includes some editing capabilities. Outside of AIPS, DIFMAP provides a highly integrated procedure that is especially useful with smaller data sets where the computations happen quickly enough that the process can be very interactive. This chapter is based mainly on the AIPS approach because of the author's experience. But for some classes of observations, DIFMAP is almost certainly a better approach, although it does not yet have robust weighting. The strongest convergence of the self-calibration procedure that the author has experienced was obtained using Dan Brigg's NNLS algorithm (Briggs 1995) which is an imaging/deconvolution algorithm based on a least squares fit. That algorithm is not available in AIPS or DIFMAP, but it is in AIPS++ and SDE. It does have some limitations on very large sources. It is especially powerful when the main emission region is partially, but not completely, resolved on the longest baselines.

Self-calibration is also done for data from linked interferometers like the VLA. With larger numbers of baselines and better starting models than is typical for VLBI, the convergence is usually faster, often 1 to a few iterations. But the very highest dynamic range images, especially on sources of some complexity, can require 20 or more iterations. The highest dynamic range images on the VLA also require special efforts to reduce and calibrate closure offsets. One important way to reduce such offsets is to set the delays more accurately than normal, either before observing or in postprocessing using a fringe fit. There is really no fundamental difference between linked interferometer and VLBI data - just differences in the degree of difficulty of certain processing steps.

6. Conclusions

VLBI has reached a new level of maturity in the last few years, largely because of the advent of the VLBA. The instrument is producing data of quality undreamed of not long ago. And by setting an example, it is pushing the rest of the astronomical VLBI world to be satisfied with nothing less. The postprocessing software has not fully kept up as much of the effort has shifted to the AIPS++ project, but it is still only somewhat harder to use for VLBI data than for connected interferometers like the VLA. This situation should continue to improve.

The high quality model used on the VLBA correlator has brought a couple of paradigm shifts to VLBI processing. There once was much concern about all the subtleties of weak source fringe fitting. Now fringe fitting can be confined to calibrators at all but the most extreme frequencies or for space VLBI. Most users should only need to face the much easier task of fringe fitting strong sources. But perhaps the most exciting development is the near routine use of phase referencing. This has meant that VLBI observations can address the science on sources of a milliJansky or less — a regime only reached in the past with heroic effort. This greatly increases the number and types of sources that can be observed.

It will not be possible to observe typical thermal sources with VLBI without increases in sensitivity of several orders of magnitude, not something that is likely to happen soon. Therefore the range of source types that can be studied is inherenetly narrower than for the linked interferometers. But the sources that can be detected can be studied with resolution far beyond what is available with any other direct imaging technique. A particular strength of VLBI, and one that will require a lot of observing time to exploit, is that the observed sources are small enough to exhibit structural changes on human time scales, even for sources at the edge of the observable universe. Perhaps more so than any other type of telescope, a VLBI array can be a movie camera.

References

Beasley, A. J. & Conway, J. E. 1995, in *Very Long Baseline Interferometry and the VLBA*, Astronomical Society of the Pacific Conference Series, Volume 82, eds. J. A. Zensus, P. J. Diamond, & P. J. Napier (San Francisco: ASP), 327–341.

Briggs, D. S. 1995, *High Fidelity Deconvolution of Moderately Resolved Sources* Ph. D. thesis, New Mexico Institute of Mining and Technology.

Carilli, C. & Holdaway, M. 1997, MMA Memo 173.

Cotton, W. D. 1995, in *Very Long Baseline Interferometry and the VLBA*, Astronomical Society of the Pacific Conference Series, Volume 82, eds. J. A. Zensus, P. J. Diamond, & P. J. Napier (San Francisco: ASP), 189–207.

Felli, M. & Spencer R. E. 1989, *Very Long Baseline Interferometry, Techniques and Applications* (Dordrecht: Kluwer Academic Publishers).

Lonsdale, C. J. & Doelman, S. S. 1998, in *IAU Colloquium 164: Radio Emission from Galactic and Extragalactic Compact Sources*, Astronomical Society of the Pacific Conference Series, Volume 144, eds J. A. Zensus, G. B. Taylor, & J. M. Wrobel, (San Francisco: ASP).

Ma, C, Arias, E. F., Eubanks, T. M., Fey, A. L., Gontier, A.-M., Jacobs, C. S., Sovers, O. J., Archinal, B. A., & Charlot, P. 1998, *AJ*, 116, 516.

Miyoshi, M., Moran, J., Herrnstein, J., Greenhill, L., Nakai, N., Diamond, P., & Inoue, M. 1995, *Nature*, 373, 127.

Napier, P. J., Bagri, D. S., Clark, B. G., Rogers, A. E. E., Romney, J. D., Thompson, A. R., & Walker, R. C. 1994, Proceedings of the IEEE, 82, No. 5, 658.

Peck, A. B. & Beasley, A. J. 1998, in *IAU Colloquium 164: Radio Emission from Galactic and Extragalactic Compact Sources*, Astronomical Society of the Pacific Conference Series, Volume 144, eds J. A. Zensus, G. B. Taylor, & J. M. Wrobel, (San Francisco: ASP).

Rogers, A. E. E., Doelman, S. S., & Moran, J. M. 1995, *AJ*, 109, 1391.

Sovers, O. J., Fanselow, J. L. & Jacobs, C. S. 1998, *Rev. Mod. Phys.*, 70, 1393.

Thompson, A. R., Moran, J. M., & Swenson, G. W. 1986, *Interferometry and Synthesis in Radio Astronomy* (New York: John Wiley & Sons).

Zensus, A., Diamond, P. J., & Napier, P. J. 1995, *Very Long Baseline Interferometry and the VLBA* Astronomical Society of the Pacific Conference Series, Volume 82 (San Francisco: ASP).

Zensus, A., Taylor, G. B., & Wrobel, J. M. 1998, *IAU Colloquium 164: Radio Emission from Galactic and Extragalactic Compact Sources*, Astronomical Society of the Pacific Conference Series, Volume 144. (San Francisco: ASP).

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

23. Astrometry and Geodesy

Ed B. Fomalont
National Radio Astronomy Observatory, Charlottesville, VA 22903, U.S.A.

Abstract. Astrometry deals with the precise measurement of the absolute and relative position of celestial objects. High quality imaging of these objects, the main concern of many of the other lectures, is less important unless the object is large compared with the resolution of the observations. Since radio measurements are made from the Earth surface, the complicated motion of the Earth in space, the distortion and slippage of the Earth surface (geodesy) and the properties of environment through which the radio waves travel, can also be measured. This lecture will describe: the measurement of the total phase delay, the fundamental astrometric equation, simple examples of the determination of astrometric and geodetic information. For VLBI observations, the measurement of the group delay, with all of its foibles is described in some detail. A typical astrometric experiment and its reduction are illustrated. Finally, there is a discussion of astrometric and geodetic endeavors at the milliarcsecond level.

1. Introduction

The background needed for astrometry and geodetic analysis of synthesis data have been given in Lectures 2 and 5. The major difference in emphasis between synthesis imaging and astrometric solutions is the following: For imaging, the nature of the calibration errors are not important as long as their effects are removed. The most common calibration method alternates observations of the target source and a nearby calibrator source (unresolved source with known position) which successfully removes all calibrator errors to first order from the target source as long as the dominant errors are well-behaved. In fact, for strong sources no calibration is necessary in order to derive an accurate image (assuming that all calibration errors are antenna-based). However, most astrometric/geodetic observations treat the radio source as a test signal of known properties and attempt to understand (model) all of the calibration contributions. In practice some combination of imaging and astrometric reductions may be necessary and the astrometric use of weak sources requires imaging.

In Section 2 we will describe the fundamental measurement (total phase delay) used for astrometry and illustrate the basic equations from which most astrometric/geodetic information can be obtained. These equations include the major calibration errors which dominate most synthesis observations: astrometric (where is the source?), geodetic (where is the telescope?), tropospheric (what's happening along the radiation path?), and temporal (what time is it?). Examples are given to illustrate the form of the calibration terms, their solution and inherent ambiguities. In this chapter we do not get into the complication of group delay and other VLBI tricks that are needed.

Because the most interesting astrometric and geodetic effects are relatively small in angular scale and cover global-sized effects, VLBI techniques are most commonly used, and this type of high resolution observations is discussed in Section 3, covering some of the ground in Lecture 22. Since many calibration errors produce phase effects which are tens of cycles in size and vary by a few cycles over minutes of time, the explicit use of the measured fringe phase in the analyses is difficult because of the inherent lobe-ambiguity. However, the group delay (derivative of phase with frequency) is an adequate substitute for the phase and the formalism in Section 2 applies, but the many problems in the measurement of the group delay are discussed.

Section 4 will be outline a typical astrometric and geodetic experiment. Results in astrometry and geodesy at the milliarcsecond level are discussed with some results presented.

2. Astrometric Fundamentals

2.1. Total Phase Delay

The natural response (without electronic modification) of a monochromatic interferometer is a quasi-sinusoidal response, $r(t)$,

$$r(t) = A\,\cos(\phi_T)\,. \tag{23–1}$$

where ϕ_T is the total phase delay of a celestial signal which has traveled in two paths (each through a different telescope) to the correlator. The total phase delay is generally a large number. For example, for a baseline (telescope separation) of 1000 km at a wavelength of 6 cm, the total phase delay is $\sim 10^7$ cycles and can change as fast as 1 kHz because of the diurnal rotation of the Earth. These signals are combined in real time or from data previously recorded on tape in a correlator. The amplitude A of the response is related to the intensity of the cosmic signal, the overall sensitivity of the telescope system, and the angular structure associated with the signal.

The largest contribution in the total phase delay is the time difference of flight between the radio source and the two telescopes. This term and other major contributers to the phase listed are:

$$\phi_T = \nu(\tau_g + \tau_n) + \phi_d + \phi_v \tag{23–2}$$

where

- ϕ_T is the total phase delay, measured in cycles; [1]
- ν is the frequency in Hz;
- τ_g is the time delay in seconds between the reception of signals at the two telescopes. It is a purely geometric quantity and is given by the famous $\mathbf{D}{\cdot}\mathbf{s}$ where $\mathbf{D}$ is the separation of the two telescopes and $\mathbf{s}$ is the direction to the source. This *geometric* delay is NOT a function of frequency since it assumes the signal travels in a vacuum;
- τ_n represents the additional contributions to the delay which *are not frequency dependent.* These are called the non-dispersive delays and are caused by the tropospheric refraction, by timing errors between the telescopes, and various path-length changes in telescope systems;

[1] In this chapter the unit of phase will be *cycles*, rather than radians. The term fringes, revolution, turns and lobes are also used in place of cycles.

- ϕ_d denotes any additional phase change which is not a linear function of frequency, i.e., not a pure time delay. These phase changes are called *dispersive* because a pulse when propagating through a dispersive medium will widen with time; and
- ϕ_v is the visibility phase which is non-zero for an extended source.

The *a priori* parameters associated with interferometric observations are generally well known, and from this information we can calculate the model phase delay, ϕ_M, the expected phase delay if only the parameters in the assumed model completely described the observation. Very important components of the model are: the source position (called the phase center), the antenna locations, and the time keeping. Other useful model components are: the atmospheric refraction; telescope path-length (cable) changes; and phase changes in the electronic systems. Some of these components are measured during the observations. It is also possible to include the effects for an extended source in the model phase delay.

The output from most correlating interferometers is the fringe phase, ϕ_f, the difference between the model phase delay and total phase delay, averaged over a period during which the fringe phase is relatively constant. [2]

$$\phi_f = \mathrm{FP}\{\phi_R = \phi_T - \phi_M + \phi_v)\} \qquad (23\text{–}3)$$

where FP denotes the fractional part. Notice that the fringe phase is not necessarily equal to the residual phase delay, ϕ_R, the difference between the true and the model phase delay, because of the lobe ambiguities of the measurement of the fringe phase. We call an interferometer phase stable, if the difference between the total phase delay and the model phase delay is generally less than one cycle and the total phase delay can be reconstructed from the fringe phase. Such phase stability is generally the case for most kilometer-sized arrays. If this model phase delay is not well-known or there are large unmodelable changes in the residual phase, then the interferometer is phase unstable. The degree of unstableness varies from modestly unstable (VLBI observations and optical interferometers) where phase stability can be achieved with a tricky calibration techniques, to fully unstable (intensity interferometer) which must use other correlation techniques to obtain usable information (e.g., Thompson, Moran & Swenson 1986, pp. 482–486).

We have gone through this relatively long description of the various phase terms used in interferometry for the following reason: Astrometric/geodetic results are derived from the total phase delay, not directly from the measured fringe phase. Thus, the array must be fairly phase stable. VLBI tricks in getting around mild instabilities are described in Section 3.

2.2. Fundamental Astrometric Formula

The following formula, taken from Lecture 1, contains the basic parameters used for nearly all of the astrometric and geodetic observations. We will start with

[2] A quantity which is called a phase is only defined between −0.5 and 0.5 cycles. Anything called a phase delay includes the number of cycles.

a few simple experiments and will assume for simplicity that the model phase delay is sufficiently accurate so that lobe-ambiguities are not too much of a problem. These types of astrometric observations have been carried out with the VLA and other kilometer-sized arrays.

The fringe phase for a point source as a function of hour angle H, defined at the midpoint of the baseline is:

$$\begin{aligned}&\phi_f = \frac{\nu}{c}\{A\ \cos H + B\ \sin H + C\}\\ &+\nu\{\tau_n + \tau_c + L_d - L_w\} + \phi_d + \phi_v\end{aligned} \tag{23–4}$$

where

$$\begin{aligned}A &= \Delta L_x \cos\delta + (\Delta\alpha - \delta t)L_y \cos\delta - \Delta\delta L_x \sin\delta\\ B &= -\Delta L_y \cos\delta + (\Delta\alpha - \delta t)L_x \cos\delta + \Delta\delta L_y \sin\delta\\ C &= \Delta L_z \sin\delta + \Delta\delta L_z \cos\delta\end{aligned}$$

where $(\Delta L_x, \Delta L_z, \Delta L_z)$ is the (true–modeled) antenna separation (equatorial components) in units of physical length. Generally, one telescope is treated as the reference telescope and the other the unknown telescope. $(\Delta\alpha, \Delta\delta)$ is the true–modeled) source position. These terms represent the first order expansion of the geometric delay error and is the 'pure' interferometric part of the residual terms. We have glibly assumed that the celestial and terrestrial coordinate systems are well-defined. However, much of the cutting edge research in astrometry and geodesy deals with the proper definition and time variations of these fundamental frames. More on this in Section 4.

We have also included a time offset, τ_c between the clocks at the two telescopes, and time offset, δt, between reference telescope time and the true time. The term τ_n represents any other non-dispersive delays. Two terms are associated with the troposphere, a serious source of error. A global tropospheric zenith delay associated with the dry troposphere, L_d, and a more variable wet tropospheric delay L_w. Finally, the non-dispersive phase terms ϕ_d and the radio structure term (visibility phase) ϕ_v are given.

The source position error and the telescope location error produce a residual fringe phase with a period of one sidereal day. The other terms have a period characteristic with time scales from seconds to years. Unfortunately, atmospheric and some temperature sensitive phase errors have a diurnal period which complicates the separation of the variables. But, first we will give some simple astrometric/geodetic examples.

2.3. Idealized Astrometric/Geodetic Solutions

The resultant fringe phases obtained from typical experiments with kilometer-sized arrays are shown in Figure 23–1. We have used a telescope baseline of 30 km and a frequency of 5 GHz (fringe size of 410 mas in the sky), and we have plotted the variation of fringe phase over a fifteen hour period for some examples. We are assuming that the phase delay model for these observations are perfect, except for the source position error and an arbitrary phase offset.

In Figure 23–1a there is an arbitrary offset of the fringe phase. If this offset is constant, then only three measurements (shown by the circles) are needed to

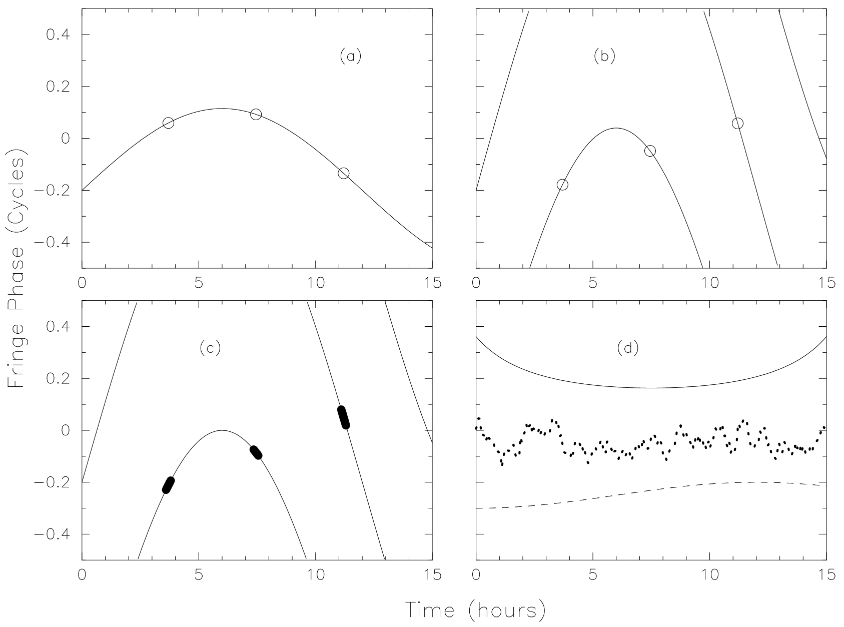


Figure 23–1. Typical Fringe Phase for a baseline of 30 km with an observing frequency of 5 GHz. **(a)**–The fringe phase for a position error of 150 mas. Three measured points are shown by the open circles. **(b)**–The fringe phase for a position error of 600 mas and three measured points. **(c)**–The measurement of the phase slope for a source displaced 600 mas. **(d)**–Typical phase errors: solid curve=residual atmosphere, dashed curve=possible thermal stretching of cable, dotted curve=short term tropospheric phase fluctuations.

determine the error in the position of the source. Of course, the further apart the three observations are made over 24 hours, the more determinate is the solution. In Figure 23–1b we have demonstrated the lobe ambiguity problem. With the three measurements indicated, an incorrect solution for the sine wave would be obtained unless there was *a priori* information about the lobe ambiguities. Such information could be obtained by making more than three observations and *connecting* the phase between successive observations; that is, adding on cycles to the measure fringe phase to produce a smooth curve with amplitude greater than one cycle. For sparse observations with many lobe ambiguities, and significant measurement errors, such phase connection is often ambiguous. Another method of analysis when there are lobe ambiguities is to measure the phase rate with time, rather than the phase, as shown in Figure 23–1c. If the phase rates are measured at only two indicated points, then the amplitude and the relative position of the sinusoid can be obtained. Notice that the C term of Eq. 23–4 drops out of the phase rate so that a clock offset and the L_z component cannot be obtained. This technique is called *fringe-rate mapping* and its algebraic form can be obtained from the time derivative of Eq. 23–4 (eg, Wohlleben, Mattes, & Krichbaum 1991).

The typical error terms for this fictitious experiment are shown in Figure 23–1d. The solid curve shows the phase error introduced by an error in the tropospheric model of 1 cm pathlength at the zenith. Since the zenith path delay at sea level is about 7 nsec (230 cm), the plotted error term comes from only 0.5% of the total expected. The large effect at low elevations (the observation is assumed centered at source transit) is obvious. The dashed curve shows the effect of a cable stretching by 3mm over the fifteen hours. If the stretching is caused by a thermal change associated with the external temperature, a period of 24 hours is reasonable. Such an error, even if small, will produce a systematic error in the determination of the source position. Finally, the dotted curve shows the typical short term fluctuations that are produced by clouds of water vapor which pass over each of the telescopes. The correlation time of the fluctuations in the plot are 3 minutes with an rms size of 6 mm. The typical delay for the 'wet' component of the tropospheric refraction is about 50 mm in a dry climate. Assuming that these errors are typical, a precision of 0.1 cycle (in this case about 40 mas) is about the best than can be expected (see Thompson et al. 1986, 407–439).

2.4. Astrometric Experiments

Since it takes only three short observations with a single baseline to determine the position of a radio source, many sources can be observed in a 24-hour period. One of the first such experiments was done with the Cambridge one-mile telescope (Elsmore and Mackay, 1969). With the accurate surveying of the telescope locations, and a very stable electronic system, the only unknowns were, in fact, the position of the radio sources.

In Figure 23–1 we assumed that the residual phase variation was caused by a source position error. The same signature would have occurred with a telescope location error and ΔL_x and ΔL_y could have been determined from the three observations. Notice that ΔL_z does not produce a change with time, only a constant phase offset. Hence, to determine L_z, the instrumental phase offset must be known. Alternatively, by observing two sources at different declination, the term $\Delta L_z \; sin \; \delta$ would have produced a difference in the phase offset for the sources. An experiment in which the position of many sources and the telescope locations were determined simultaneously was made by Wade (1970) using the NRAO 3-element interferometer at Green Bank. The analysis technique is derived from Eq. 23–4.

There are also ambiguities and properties in Eq. 23–4 which should be understood. First, absolute declinations and the L_z component can be determined from the variation of the C-term with declination. However, an equivalent east-west rotation can be made in three different ways: change the origin of right ascension; rotate the telescope locations around the North Pole; offset the clock at the reference telescope. This ambiguity is removed by fixing two of the three parameters; such as the origin of right ascension, or the absolute location of one telescope.

For one-baseline experiments (which were essentially the Cambridge and Green Bank experiments), we can analyze the redundancy in the solutions. For N radio source sources, at least 3N observations of each must be made in order to determine a position for each source, as described above. These can be short so that the number of sources that can be observed in 24 hours can be large. The total number of unknowns are 2N-1 (N declinations and N-1 right ascensions) +

3 baseline component errors + 1 phase offset [3] As long as N is greater than 4, a solution is possible. In practice, five or six observations are made for a source and other important parameters can also be determined; for example the zenith path delay in four hour intervals. As long as this interval contains a sufficient number of observations of many sources at a range of zenith angles, a realistic solution of the zenith path delay can be obtained. However, as with large least-square methods involving numerous coupled variables, one must be very careful in getting robust solutions. We will discuss this multi-parameter solutions more with the VLBI program CALC/SOLVE.

Finally, multi-element two-dimensional arrays enable accurate positions and baselines to be determined even with only one observation of a source. For example with the VLA, 26 independent measurements are made at any time (not 351, the number of baselines since closure is assumed). Thus, observation of N sources will give 26N measurements. The number of unknowns are: 26 telescope phase offsets, 2N-1 source coordinates and 26x3 telescope location offsets—for a total of 26x4 + 2N - 1 unknowns. Again, it is possible to include atmospheric properties or a piecewise linear set of phase offset. Although only five or six observations need be made, most astrometric programs at the VLA are 12 to 24 hours in length, observe over 100 sources in about 300 observations. Most sources are observed several times, although one observation of a source is enough to obtain its position.

In practice the above purely astrometric scheme (where no external reference sources are used) is rarely used at the VLA. The *a priori* telescope locations are well-known and there is a grid of hundreds of unresolved sources (calibrators) which have known positions with an accuracy of 10 mas. Most programs alternate observations of the unknown sources with nearby calibrators. To first order all phase errors (caused by small antenna location errors, by clock errors, by atmospheric refraction) are removed. The source positions can be determined from mapping or from model fitting of the source fringe phase from all baselines, after the above calibration with a nearby calibrator. Accuracies of 50 mas are generally obtained in A-configuration at 8 GHz using this calibration method. However, to reach accuracies better than 30 mas, pray for good weather and use the astrometric scheme described in the above paragraph.

3. VLBI and the Group Delay

In the previous section we discussed astrometric observations using kilometer-sized arrays where lobe ambiguities of the measured fringe phase were not much of a problem. However, these small arrays had accuracies limited to about 30 mas. Much higher angular precision at about the one milliarcsecond level can only be reached with VLBI using global arrays with baselines 100 times longer. These global arrays are also much more geologically interesting.

[3] This instrumental term can be generalized if necessary. If the instrumental phase changes with time, one could determine a phase offset for any four hour interval for example In this case there would be 6 phase offsets to solve for in 24 hours.

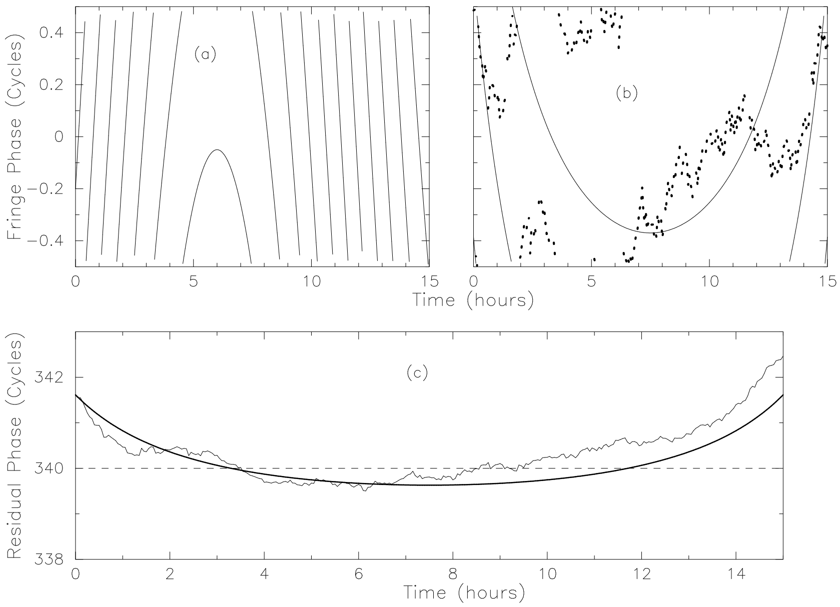


Figure 23–2. Typical Phase Variations for VLBI baselines of 5000 km and an observing frequency of 5 GHz. **(a)** The fringe phase for a position error of 150 mas. **(b)** The fringe phase error caused by – solid line, a tropospheric zenith path error of about 3 cm from the assumed model; and – dotted line, typical clock noise and tropospheric fluctuations. **(c)** The residual phase delay produced by – dashed line, a 68 nsec clock offset, – heavy line, the residual phase associated with the position offset only; and – light line, the sum of all phase terms present in the data.

3.1. Group Delay and Lobe Ambiguities

All of the fringe phase plots in Figure 23–1, with VLBI resolutions, scale by a factor of 100, and thus spread over many cycles. A typical example for a baseline of 5000 km and a frequency of 5 GHz is shown in Figure 23–2. The astrometric analysis using the fringe phase implied by Eq. 23–4 would be almost impossible to use with these data. Even the phase change with time is too contaminated a measure, because of short term variations, to produce accurate results.

The residual phase delay contains unmodeled tropospheric phase errors of many cycles, especially at low elevations. Short term fluctuations of period minutes to hours are also associated with the clock difference between and atmosphere over the two telescopes. A large residual phase offset is caused by the clock difference between the two telescopes. However, even with these relatively large errors in the residual phase delay, the 24-hour sinusoid caused by the position error is still dominant. But, this behavior is completely lost in the measured fringe phase because of the lobe ambiguities.

The way out of this mess is to measure the rate of change of phase with frequency, called the *group delay.* This delay is then proportional to the residual phase delay **because all of the major phase terms are associated with true delays and are linear with frequency.** Thus, by measuring the fringe phase at many frequencies simultaneously, one can obtain the group delay (with complications we will get into in a moment). All of the discussion associated with Eq. 23–4 will still apply since the previous discussion is relevant with group delays as it is with fringe phase. More formally, the group delay, τ_{gr}, is defined as

$$\tau_{gr} = \frac{d\phi_f}{d\nu} = \tau_g + \tau_n + \frac{d}{d\nu}(\phi_d + \phi_v) \qquad (23\text{–}5)$$

3.2. Measuring the Group Delay and Bandwidth Synthesis

The measurement of the group delay is not as simple as it appears because

- The relative accuracy for determining astrometric parameters in using the group delay, compared with the phase delay, is about equal to $\Delta\nu/\nu$ where $\Delta\nu$ is the frequency range used to determine the group delay. For a bandwidth of 128 MHz at 5 GHz, the group delay is only about 4% as accurate as the phase delay.

- Some of the frequency channels must be sufficiently close so that lobe ambiguities are not a problem in determining the phase slope. For example, if the group delay is 1000 nsec, then at least two frequencies must be closer than 500 kHz to define an ambiguous slope. A safety margin of at least a factor of five is useful here.

In the next ten years when recording media become more dense and correlators can handle a GHz bandwidth, group delay determinations will be straightforward. However, at the present time an instantaneous bandwidth of 128 MHz is the usual limit.

The technique of bandwidth synthesis (Rogers 1970) is used to determine unambiguous and accurate group delays even with relatively narrow bandwidth systems. Most correlators naturally split the incoming radio signal into narrow channels (see Lecture 4) so the ambiguity problem is fulfilled. Since the 1970's most astrometric instruments have been designed to observe at many independent frequencies which need not be contiguous. Hence, even if only 64 MHz of instantaneous bandwidth is available this bandwidth can be chopped up into, for example, 8 frequency IF's (as they are called at the VLBA), each of 8 MHz width, which span as large a bandwidth as possible, perhaps as much as 1 GHz at 5 GHz observing frequency. The group delay will then be much more accurately determined.

A typical frequency configuration of a VLBA observation at 1.5 GHz, designed for astrometric observation of pulsars, is shown in Figure 23–3. There are two levels of frequency segregation. First, a relatively wide-band signal (1.4 to 1.7 GHz) can be detected and amplified by the telescope feed and front-end electronics. Since this wide bandwidth signal cannot be handled with present correlators, it is split into several (typically 4 to 8, although the MKIII system has 14) discrete, relatively narrow-band IF paths. Each IF bandwidth is typical 4 to 16 MHz, but the frequency of the IF's are spread as much as possible. Simultaneous observations at two frequency bands which are well-separated are particularly useful for removing the ionospheric component.

Each IF is correlated independently and the output fringe phase is determined at many equally-spaced frequency points throughout each IF. The number of channels in each IF can be as large as 8092 in some correlators; however, the

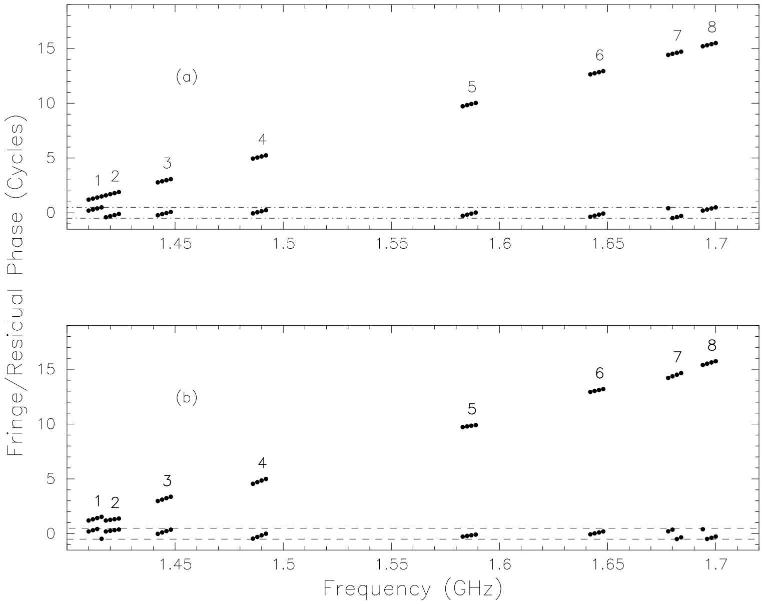


Figure 23–3. Phase versus Frequency: **(a)** An ideal observations. The fringe phase is measured in four channels for each of the eight IF's, spread from 1.41 GHz to 1.68 GHz. The fringe phase is restricted to −0.5 to 0.5 cycles, and corresponds to the plotted residual phase delay. From the closely spaced frequencies, the approximate group delay can be estimated and then extended to the multi-cycle residual delay without ambiguity. This delay value is 49.3 nsec. **(b)** A typical observation. In reality, each IF has an arbitrary phase offset and an additional slope. These are caused by the peculiarities in the IF circuitry. Without the removal of the slope and, especially, the phase offset, it is difficult to determine the group delay from the fringe phase.

number of channels is usually averaged to 8 to 32 for most applications. Thus, the minimum separation of frequency channels can be as small as 100 KHz, or even smaller if necessary.

In Figure 23–3a both the residual phase delay (over many cycles) and the measured fringe phases (only defined between −0.5 and 0.5 cycles) are shown. Although the fringe phase is sampled every few seconds, the phase difference between all of the frequency channels remains relatively constant over minutes of time and the data can be averaged, using the technique of fringe-fitting. The phase slope over the individual IF's is called the Single Band Delay (they should be equal for all of the IF's) and the phase slope defined, collectively, by the eight IF's is called the Multi Band Delay, as well as the group delay. The choice of which IF frequencies to pick can be crucial in order to span as much frequency as possible without the possibility of ambiguities.

3.3. Complications in Measuring the Group Delay

The measurement of the fringe phase and residual phase delay for a typical VLBI observation for the same observation is shown in Fig. 23–3b. Two complications have been added: The phase offset among the individual IF's, and the peculiar single-band delay and group delay offsets. Notice that with the actual fringe phases measured at the bottom of Figure 23–3b the group delay could not be determined because of the effect of the phase ambiguities.

Hence, the phase offset for each IF must be determined before the group delay can be determined. Some arrays use multi-frequency test signals to measure these offsets directly (they can vary with time and with other observing parameters like source elevation) and these measurements and their application are discussed in Lecture 22. They can also be determined in an ad-hoc manner in the following way—called *The Manual Phase Cal.* Choose one short observation of the good quality source (unresolved source, with an accurate position, with good signal to noise and complete IF coverage) and fringe-fit the data to each IF separately. The result will be a single-band delay and a phase offset for each of the IF's. [4] Apply the IF single-band delays and phase offset to the entire experiment. By definition, the short observation will now have zero single-band delay and zero-multi-band delay. If the peculiar single band delays and the IF phase offsets are constant with time (this is not always a good assumption), the fringe phases for all of the other sources will be linear with frequency across all of the IF's. If these residuals are not constant with time, non-linearities of the phase with frequency will occur at other times and the group delay determination will be in error, or impossible to obtain if the phase changes are more than about 0.1 cycle (36°). Such deviations can and should be checked by looking at the group delay determination for other good quality sources throughout the observations.

The major non-dispersive delay/phase effect is that caused by the ionosphere. If I is the ionospheric phase contribution at some reference frequency ν_0, then the ionospheric contribution at any other frequency ν is $I(\nu_0/\nu)$, and is a relatively large contribution at frequencies less than 3 GHz for VLBI baselines. Short-term ionospheric changes often occur just before sunrise and just after sunset and the more constant ionospheric refraction is difficult to model. Typical ionospheric phase contributions at 1.5 GHz are about 10 nsec. At 8 GHz it is about 0.2 nsec, more than one cycle. The technique of removing the ionospheric component of the group delay, using simultaneous dual frequency observations, is illustrated in Figure 23–4 and the algorithm is

$$\tau = \frac{\nu_0^2 \tau_0 - \nu^2 \tau}{\nu_0^2 - \nu^2} \tag{23–6}$$

Virtually none of the good quality radio sources are point sources with VLBI resolution, and many have low correlated flux densities at the longest spacing. Their visibility phase contribution, ϕ_v, spoils the linear relationship of fringe phase with frequency and this error, along with the unknown atmospheric refraction, is one of the bigger contributions to the astrometric solutions. The effect of the visibility phase of the source can be determined by first imaging the source, using self-calibration techniques. The image can then be used to calculate the visibility phase of the source for any observation at any baseline. Several methods are available but all of them essentially remove the visibility

[4] The linear drift of phase with time will also be calculated from the fringe fit. This phase rate term is not applied, but permits the averaging of the data over a longer period of time.

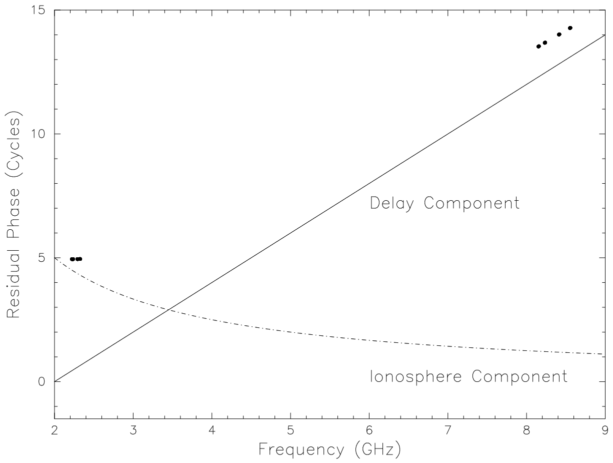


Figure 23–4. Dual Frequency Group Delay: The non-linear Ionospheric phase contribution and the non-dispersive phase contribution can be separated by measuring the group delay at two different frequencies. The delay from the four S-Band frequencies (2.20, 2.21, 2.29, 2.32 GHz) is 0.025 nsec. The delay from the four X-Band frequencies (8.15, 8.23, 8.41, 8.55 GHz) is 1.855 nsec. The corrected non-dispersive delay component is 2.000 nsec.

phase term based on the source structure. Since the structure of radio sources changes significantly between 2.6 and 8.1 GHz, for example, the removal of the ionospheric contribution is also sensitive to source structure changes.

The choice of IF frequencies must balance the need for close spacings to avoid ambiguities and the need for wide spacings for maximum spanned frequency. However, in relatively low signal to noise observations, it is possible to obtain a group delay which is grossly in error. This ambiguity is illustrated in Figure 23–5a and b for the frequencies 8.15, 8.23, 8.41 and 8.55 GHz. The ambiguity every 50 nsec is produced by the largest common multiple of 0.02 GHz for the four frequencies. Often, this ambiguity can be resolved using the single band delay value obtained from the individual frequencies sampled in each IF. Figure 23–5b shows why there is a relatively good chance of picking a delay which is in error by 6.5 nsec. These mini-ambiguities are common when the number of IF's used is less than about five with low signal to noise observations.

4. VLBI Astrometric/Geodetic Results

4.1. Observing and Reducing a 24-hour Experiment

A typical astrometric/geodetic experiment lasts for 24 hours, in which about 300 separate observations are made. From 30 to 50 sources are observed with good mixing of elevation, azimuth and sources. Observations over different parts of the sky are mixed together and perhaps half of the observation interval is used slewing from one source to another. This type of observing is needed for determining the tropospheric refraction which is large and variable. It can only be determined with many observations at varied elevations over a few hours. This

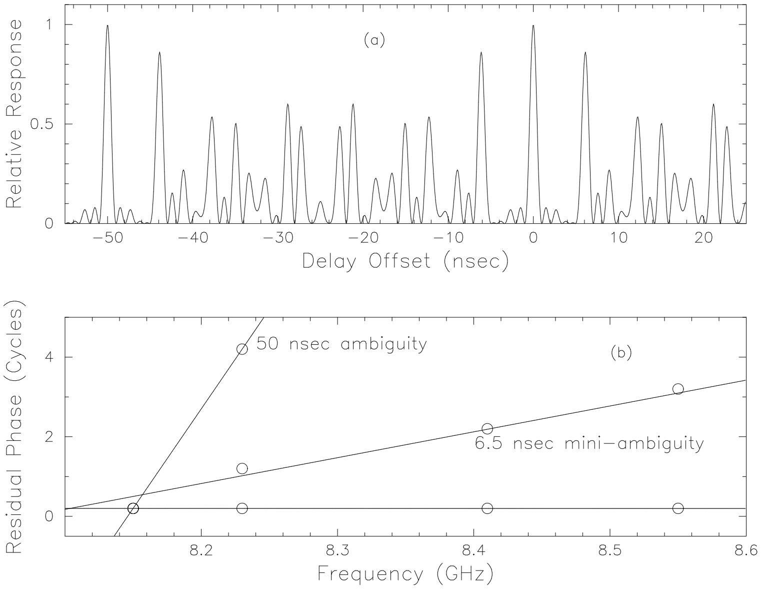


Figure 23–5. Group Delay Ambiguities: **(a):** The delay function shows the relative probability of obtaining a delay offset from the true delay. The unity peaks at multiples of 50 nsec are the main ambiguity spacing. The peaks at 0.85 level are mini-ambiguities at 6.5 nsec separation and observations with less than 7:1 signal to noise may fit the data at this offset. **(b):** Three fits to the same fringe phase measurement of 0.2 cycles at the four frequencies. The fits at 0 nsec and 50 nsec (last two points off scale) are perfect. The fit at 6.5 nsec is possible with phase errors of about 0.05 cycle.

mode of observation is identical for a phase connected, short-baseline experiment or for a group-delay, VLBI experiment.

After correlation of the data, the residual phase and delay terms are calculated using the AIPS or other interferometric package. After determining the appropriate phase calibrations in order to linearize the phase versus frequency for each IF, a fringe-fitting algorithm (see Lecture 22) determines the phase and phase slopes. These quantities are averaged for about five minutes of data and referenced to a fiducial time near the middle of the observation. The **total** phase delay, group delay, and delay rate for each baseline are calculated. Other measurements which can be used in the subsequent fitting packages are also supplied; e.g., meteorological information at each site. These data are then transferred to one of the software packages which can solve for the astrometric/geodetic parameters. Two systems in common use are NASA Goddard's CALC/SOLVE package (eg NASA, 1980) and NASA JPL's MODEST (Sovers & Jacobs 1994).

A description of using the CALC/SOLVE software has been given by Shaffer (1993) and will not be repeated here. Plots of residual delay with time (using CALC/SOLVE), illustrating the basic steps are shown in Figure 23–6. CALC first determines the residual delays from the input total delays and a model delay which is resident in CAL. Partial derivatives are also calculated for the SOLVE part. The largest residual delays are caused by clock drift between the telescopes and by inaccurate modeling of the tropospheric refraction. When these are estimated (in two hour blocks), the residual delay remaining in Fig. 23–

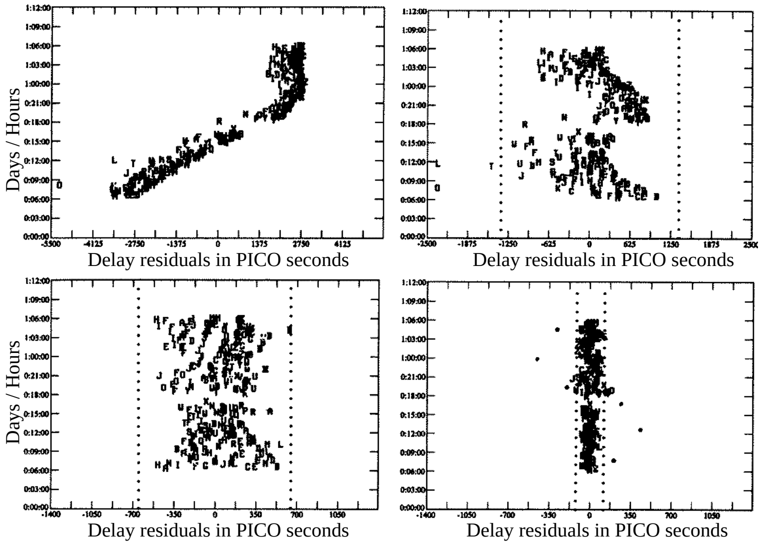


Figure 23–6. Solution Steps using CALC/SOLVE which show the residual delays from the Kitt Peak to Fort Davis Baseline at 8 GHz: **(a)** The calculated residual delays from the measured total group delays minus the accurate model calculated in CALC. The clock variation is about 5 nsec, with an offset of 955 nsec. **(b)** The residual delays after fitting for a clock offset, drift and acceleration during the day. The points spread to the left are caused an inaccurate tropospheric delay model mainly affecting by low elevation observations. **(c)** The residuals delays after fitting 2-hour piece-meal clock offsets and atmospheric zenith path delay terms. **(d)** The residual delays after determining the best fit source positions and telescope locations, and nutation terms.

6c are nearly all dominated by errors in the geometric delay. After solving for the source positions and antenna locations, the residual delay in Figure 23–6d has dropped to about 40 psec, equivalent to about 0.3 cycle at 8 GHz. This is a typical residual.

The uncertainties in the tropospheric model are believed to be the dominant error contribution after obtaining the best fitting solution. Many improved models have been proposed and investigated over the last decade; some models use the meteorological measurements, others use average weather conditions at each site. Those interested in the gory details can look in the following references (Chao 1974, Lanyi 1984, Davis et al. 1985, Herring 1992, Niell 1994).

4.2. Astrometry and Geodesy at the 1 mas level

At the present time, the accuracy attained by these 24-hour experiments is about 0.3 mas in angular scale, about 6 mm in terrestrial scale. At this level of accuracy, there are many effects which are not explicitly mentioned in connection with Eq. 23–4, but are at the cutting edge of astrometry and geodesy. Only a brief description can be given here.

From a typical 24-hour experiment, improved radio source positions and telescope locations are obtained. The tropospheric and clock parameters, though not of astrometric use, must be accurately determined. With the typical sensitivity of these experiments, variations in nutation, polar motion and UT1 should also be determined (see below for more details) since these effects cannot be predicted with mas accuracy. These results and then stored in a compact format and software in CALC/SOLVE is available to determine the changes of parameters (for example the drift of the location of an particular telescope) over years and decades. There are over 2000 such data bases and one million observations available in this global data base.

Celestial Reference Frame: The quasi-inertial frame used for astronomical observations is tied to the radio sources and quasars which are assumed very distant and fixed in the sky. The origin of the frame is the solar system barycenter (SSB). Although using the center of the Milky way or removing the motion implied by the dipole term of the cosmic microwave background radiation would be better, systematic errors using the SSB are less than 1 μsec. The system now in use is the J2000.0 system which defines the direction of the fundamental coordinate axes at a specified time. The transformation of terrestrial observations to the SSB requires accurate knowledge of the earth orbital motion (Standish and Williams, p 173, IAU 141). The best data come from radar ranging techniques (DE202/DE200 ephemeris) and the SSB is known to an accuracy of 100 m or 3 mas. The positions of pulsars determined by pulse timing methods also require the accurate Earth motion and the effects of the planets on the propagation of the radio waves in the solar system. The comparison of interferometric and timing positions of pulsars are an important check on the consistency of the reference frame definitions (Backer et al. 1985, Fomalont et al. 1992).

Radio Source Coordinates: Many VLB groups have determined accurate radio positions of small-diameter radio sources. The most complete description and results are given by Ma et al. (1990). Up-to-date compendia of radio positions and telescope locations are kept by Eubanks (1998) and are available in the NRAO VLBA data base. As part of a campaign to find small-diameter radio sources and to measure their accurate positions, the USNO is observing several times a year with the VLBA (Fey 1998). Results are on the USNO web site. Positional accuracies are now well below 1 mas for high quality sources. Because most sources have structure at the milli-arcsecond level, a radio grid accuracy of 0.5 mas or less will require the continual measurement of many radio sources in the sky to average out the structure changes.

Earth Orientation: The largest change in the direction of the rotation axis of the pole is the luni-solar precession with an amplitude of 23.5° with a period of 26,000 years. There are smaller terms associated with the planets and a relativistic effects. Shorter time-scale precession, produced by a complicated earth-moon interaction is called *nutation*. It has a period of 18.62 years and an amplitude of $9''$. Recent VLBI observations have found components of nutation with a period of one year and 13.7 days. These are sufficiently difficult to predict accurately that nutation is generally included as one of the unknowns in a standard solution for a 24-hour observation.

Earth Rotation: The phase of the earth rotation is also difficult to predict more accurately than about a few milliseconds of time. This offset in the assumed and measured UT1 is also included in the 24-hour experiments (the δt term). The length of day is decreasing 1 to 2 msec per century, and there are irregular changes of 4-5 msec over decades and 1 msec variations with periods of weeks. Changes in the length of day are also associated with storms which exchange angular momentum between the earth and the atmosphere, and the El Nino current. NASA Goddard has be recently begun a project to measure hourly variations of the Earth rotation which are occurring. The cause is unknown.

The time which is tied to the rotation of the earth is called UT1 and it defines a global sidereal time. (UT0 defines the local sidereal time and differs as much as 35 msec from UT1 because of polar motion, see below). As described above, UT1 does not flow uniformly. International Atomic Time (IAT), is based on the frequency of radiation of a transition of the cesium 133 atom. A derived time, called Universal Time, UTC, runs at the same rate at IAT, but an offset is periodically added to UTC so that the difference between (UT1-UTC) remains less than one second. Most time services distribute UTC, and UT1-UTC are measured and extrapolated by the IERS Bulletin (e.g., McCarthy 1992). There is also a proper time associated with the SSB, called TDB.

Polar Motion: The intersection of the Earth rotation axis with the earth crust changes with time and this effect is called polar motion. The motion has a period of about one year (driven by the atmosphere) and another period of 433 days (Chandler Wobble) of unknown origin. Both changes have an amplitude of about $0.15''$, or about 9 meters. Measurements and predictions of polar motion are also supplied by the IERS. These terms are included in many 24-hour experiments since the predictions are less accurate than the experiment sensitivities.

Terrestrial Reference Frame: For a discussion about the terrestrial reference frame, see *Astronomical Almanac*, K11. The IERS Terrestrial Reference FRAME (ITRF) publishes a list of about 200 reference sites around the world, mostly VLBI, GPS and SRL stations. This grid is accurate to about 10 cm. corresponding to 5 mas on the sky.

Earth Deformations: Measuring Earth deformations is one of the geodetic goals of VLBI observations. Most of these changes are global in nature and require long baselines for that reason. Some deformations are local to the telescope (where is the true focal point and is the telescope sagging with time?) and can be larger than 1 cm. The measurements of plate rotational velocities on the Earth using geological theories (e.g., Argus & Gordon 1991) and VLBI observations are in excellent agreement. Of particular importance are vertical motions which are less well understood. These changes, unfortunately, are strongly coupled to the tropospheric refraction models and have larger uncertainties than horizontal motions. Other crustal effects for which VLBI are sensitive are: Solid Earth Tides (e.g., Cartwright & Edden 1973); Ocean loading (effect of ocean tides on coastal regions) (e.g., Scherneck 1991); atmosphere loading (e.g., Rabbel & Schuh 1986).

4.3. Future Research

Over the next decade, the following astrometric and geodetic research is needed to reach the 1 mm and the 0.1 mas level. For astrometric improvement, radio source structure must be determined and used in the determination of the group delay. The goal is to produce a catalog of radio sources which cover the sky with an accuracy of 0.1 mas. Tropospheric modeling can be improved, perhaps with atmospheric ray tracing algorithms. Methods of measuring the wet component from the water vapor emission properties in the line of sight may be useful. Antenna deformations at the mm-level are also significant.

The accuracy of GPS (Global Positioning System) measurements of ground stations has passed that obtained by VLBI techniques–mainly due to the large number of satellites and unlimited observational time. However, VLBI techniques are still crucial for several reasons. The terrestrial reference frame can only be connected to the fundamental quasar reference frame by extensive VLBI observations, at least once per month. Short-term, hourly changes in earth rotation and other earth instabilities can be measured more accurately using VLBI techniques. Finally, the vertical motions in the earth crust and those associated with the many tide-like crustal deformations are limited by the tropospheric refraction variations and affect both the GPS and VLBI observations. At the level of several mm, there are differences between the theoretical geological models and measurements or the vertical component.

References

Argus, D. F. & Gordon, R. G. 1991, *Geophys. Res. Let.*, 18, 2039–2042.

Backer, D. D., Fomalont, E. B., Goss, W. M., Taylor, J. H. & Weisberg, J. M. 1985, *AJ*, 90, 1275.

Cartwright, D. E. & Edden, A. C. 1973, *Geophys. J. Roy. Astron. Soc.*, 33, 254–264.

Chao, C. C., 1974, *The Troposphere Calibration Model for Mariner Mars 1971*, Technical Report 32-1587, JPL, Pasadena, 61–76.

Davis J. L, Herring, T. A., Shapiro, I. I., Rogers, A. E. E. & Elgered. G. 1985, *Radio Science*, 20, 1593–1607.

Elsmore, B. & Mackay, C. D. 1969, *MNRAS*, 146, 361.

Eubanks, T. M. 1998, private communication.

Fey, A. 1998, private communication.

Fomalont, E. B., Goss, W. M., Lyne, A. G., Manchester, R. N., & Justtanont, K. 1992, *MNRAS*, 258, 497.

Herring, T. A. 1992, *Refraction of Trans-atmospheric Signals in Geodesy*, eds. J. C. DeMunck & T. A. Th. Spoelstra, Delft, Netherlands.

Lanyi, G. E. 1984, *Telecommunications and Data Acquisition Prog. Rept. 42-78*, 152–159, JPL, Pasadena,

Ma, C., Shaffer, D. B., de Vegt, C., Johnston, K. J. & Russel, J. L. 1990, AJ, 99. 1248.

McCarthy, D. D. 1992, International Earth Rotation Service, *IERS Technical Note*, Paris, France.

NASA 1980, *Radio Interferometry Techniques for Geodesy*, 2115 in NASA Conf. Publ., (NASA: Washington).

Niell, A. E., 1994, private communication.

Rabbel, W. & Schuh, H. 1986, *J. Geophysics*, 59, 164–170.

Rogers, A. E. E. 1970, *Radio Sci.*, 5, 1239–1247.

Scherneck, H. G. 1991, *Geophys. J. Int.*, 106, 677–694.

Shaffer, D. B. 1995, *ASP Conference Series*, 92, 345.

Sovers, O. J. & Jacobs, C. S. 1994, JPL Publication 83-39, Rev 5.

Thompson, A. R., Moran, J. M. & Swenson, G. W. Jr. 1986, *Interferometry and Synthesis in Radio Astronomy*, (Wiley: New York).

Wohlleben, R., Mattes, H. & Krichbaum, T. 1991, *Interferometry in Radioastronomy and Radar Techniques*, (Kluwer: Dordrecht)

Wade. C. M. 1970, *ApJ*, 162, 381.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

24. Spectral Line VLBI

Mark J. Reid
Harvard-Smithsonian Center for Astrophysics
Cambridge, MA 02138, U.S.A.

Abstract. Spectral line VLBI observations of cosmic sources of maser emission have contributed some spectacular astronomical results. This chapter covers basic concepts of spectral line VLBI data analysis.

1. Spectral Line VLBI Sources

A variety of astronomical sources can be observed with Very Long Baseline Interferometric (VLBI) techniques. The primary observational limitation is that, for a given VLBI array and integration time, there is a minimum brightness temperature for detection. Since, for line observations, bandwidths are limited to the line width of the source (usually less than 100 kHz), a characteristic minimum brightness temperature for the VLBA will be $\sim 10^9$ K. This high brightness temperature requires non-thermal or coherent emission processes.

High-brightness spectral lines are achieved in cosmic sources through stimulated emission (masers) and by normal (thermal) absorption of a bright background source (e. g., a compact extragalactic source). Although some effort has been made to observe 21-cm hydrogen absorption against active galactic nuclei, most line-VLBI observations to date involve masers, and this chapter will be limited to these observations. However, the techniques of absorption-line observations are similar to those of emission-line observations, and most of the principles discussed below apply to both types of sources.

Briefly, cosmic sources of maser emission are divided into three classes: interstellar and stellar masers and those associated with galactic nuclei. Interstellar maser sources are found in star forming regions. The masing molecules appear to be remnant interstellar material that did not get incorporated into stars. Widespread and strong interstellar masers include hydroxyl (OH), water vapor (H_2O), and methyl alcohol (CH_3OH). Stellar masers occur in circumstellar envelopes ejected from red giant and super-giant stars. Thus, the masing molecules are from the star. In addition to hydroxyl (OH) and water vapor (H_2O), stellar masers exhibit strong silicon monoxide (SiO) masers. Galactic nuclei display so-called mega-masers involving OH and H_2O molecules.

Most interstellar and stellar masers are associated with highly luminous ($L \sim 10^4\ L_\odot$) objects. A maser source is often composed of dozens of "spots", each at a distinct radial velocity. Spot sizes are usually less than ~ 1 AU and the spots are spread over regions up to about $\sim 10^4$ AU. Each spot demarcates a path through an interstellar cloudlet or section of a circumstellar envelope that is roughly 1 to 10^2 AU long and contains a mass of molecular hydrogen of roughly ten times the mass of Jupiter.

Since maser spots are very bright, they are often easily detected with VLBI arrays. Maps of the locations and radial velocities of masers can be used to understand the dynamics and physical conditions in the masing regions. Also, proper motions (motions on the plane of the sky) have been measured for interstellar OH and H_2O masers. Thus, the three-dimensional velocity field can be measured, and distances can be directly determined by comparison of the angular (proper) motions and the linear (radial) speeds.

2. Basic Concepts

The output of an interferometer correlator can be represented as complex numbers (fringe visibilities) on a two-dimensional grid. For a conventional delay-lag (XF) correlator the output is the complex visibility as a function of delay, τ, and time, t. The complex visibility, $V(t,\tau)$, can be represented as

$$V(t_k,\tau_\ell) = c(t_k,\tau_\ell) + i\ s(t_k,\tau_\ell)\ , \tag{24–1}$$

where $c(t_k,\tau_\ell)$ and $s(t_k,\tau_\ell)$ are the "cosine channel" or real part and the "sine channel" or imaginary part of the visibility, respectively. The subscripts denote that these variables, in practice, are not continuously determined, but are measured at discrete values. After correlation the delay function (at any point in time) can be Fourier Transformed to yield a source spectrum. For an FX correlator, such as the VLBA correlator, a short time-series of voltages from two telescopes are first Fourier Transformed and then cross-correlated. The output is then the complex visibility as a function of frequency (or Doppler velocity) and time. For either correlator, the *time sequence* of complex visibilities can be Fourier Transformed to yield the fringe frequency or fringe rate. Thus, using Fourier Transforms in the delay and/or (post-correlation) time domain, the complex visibility data can be thought of occupying four "data planes".

The complex fringe visibilities for a continuum and spectral line source are schematically illustrated in Figures 24–1 and 24–2, respectively, on the four data planes. Along each axis we show typical interferometric responses. The interferometric visibilities are complex numbers. These are difficult to portray graphically, and, with the exception of the time axis, we show only the magnitude (or amplitude) of the visibility along each axis. For the time axis we indicate the variations of the real and imaginary components with solid and dashed lines, respectively. As mentioned above, a conventional XF correlator yields visibilities in the delay–time domain (plane "a") and the VLBA FX correlator yields visibilities in the frequency–time domain (plane "c"). Given the data in any one of the 4 planes, one can manipulate, via Fourier transforms, the data to any of the remaining 3 planes as needed to simplify the analysis. The detection of a signal is usually most easily accomplished when most of the power is concentrated in a few pixels of one of the data planes.

Consider, for example, the problem of detecting a continuum source (e. g., in Figure 24–1) with an interferometer. Such a source fills the receiving band and, as such, has a narrow delay function. For a rectangular bandpass, the real part of the delay function is proportional to the Fourier transform of the rectangular bandpass given by

$$c(\tau_\ell) \propto \frac{\sin(2\pi B\tau_\ell)}{(2\pi B\tau_\ell)}\ , \tag{24–2}$$

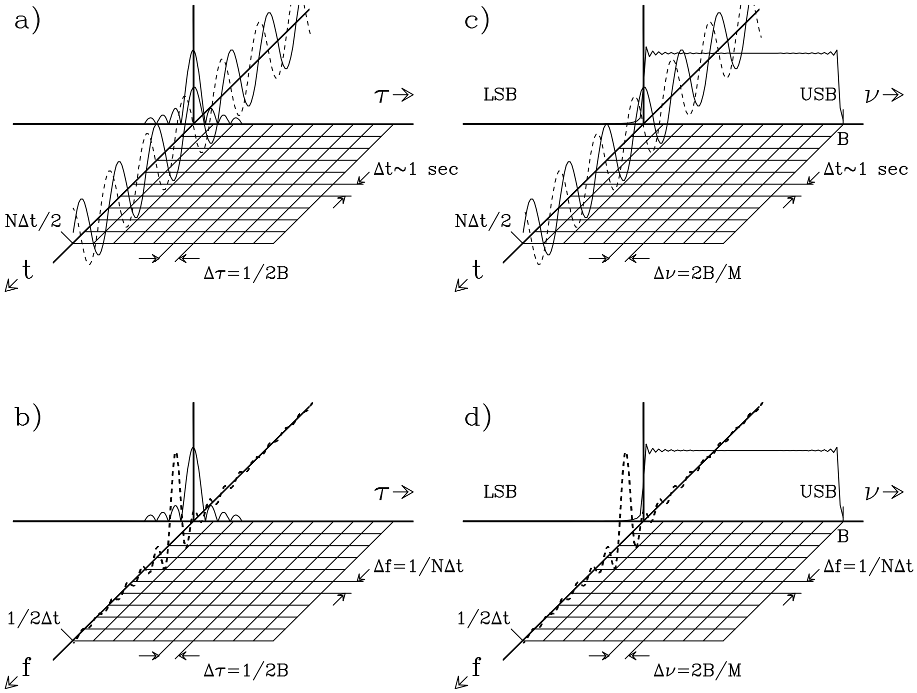


Figure 24–1. Fringe visibilities for a continuum source *(see text).*

where B is the frequency width of the bandpass and τ_ℓ is the delay value. The imaginary part of the delay function is given by

$$s(\tau_\ell) \propto \frac{\sin^2(\pi B \tau_\ell)}{(\pi B \tau_\ell)} \quad . \tag{24–3}$$

For delay "channel" spacings of $\frac{1}{2B,}$ there are two channels between nulls in the real or imaginary portions of the complex fringe visibility. It can be easily shown that the magnitude (amplitude) of the delay function (i. e., $\sqrt{c^2 + s^2}$) is $\sin(\pi B \tau_\ell) \;/\; (\pi B \tau_\ell)$ with a width between first nulls of about 4 delay channels. See Thompson, Moran, & Swenson (1986, p. 254) for more details.

In the time domain, the fringe visibility oscillates sinusoidally at the residual fringe rate (i. e., the difference between the source fringe rate and the processor or modeled fringe rate). If one Fourier transforms the time series of visibilities in each delay channel, one replaces the time axis of plane "a" with the fringe rate axis of plane "b". The fringe rate function ideally is a delta function and usually is quite narrow, provided that the time series of visibilities is shorter than the coherence time of the interferometer. Thus, in plane "b", the fringe visibility amplitude for a continuum source peaks sharply in both axes and is most easily detected there. Fringe fitting of continuum sources, therefore, is usually done in the delay–fringe rate plane.

Alternatively, the narrow-band nature of a spectral line source suggests a different approach for detection with an interferometer. Such a source has its

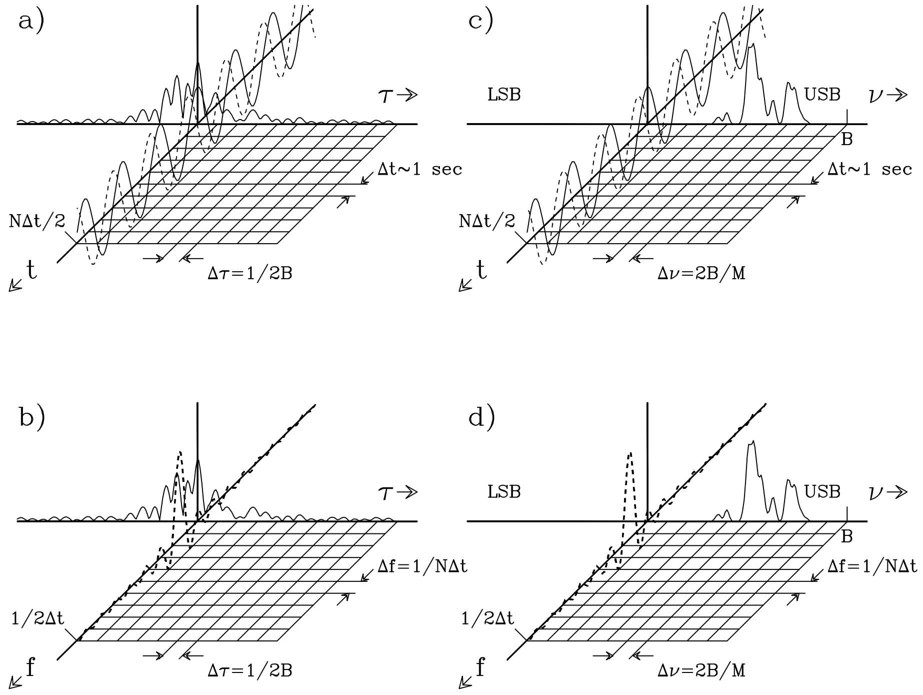


Figure 24–2. Fringe visibilities for a spectral line source *(see text).*

power naturally concentrated in the frequency domain. Since the Fourier transform of a narrow function is broad, a source with a narrow frequency spectrum has a *broad* delay function. The spectral-line source shown in Figure 24–2 is a blend of about 20 narrow components spread over about 1 MHz bandwidth, and hence has narrower delay function than would be found for any one of the components individually. In the time domain, the fringe visibility oscillates sinusoidally, as it does for a continuum source, and if one Fourier transforms the time series of visibilities in each frequency (spectral) channel, one replaces the time axis of plane "c" with the fringe rate axis of plane "d". Thus, in plane "d", the fringe visibility amplitude of a narrow line source peaks sharply in both axes and is most easily detected there. Fringe fitting of spectral line sources, therefore, is usually done in the frequency–fringe rate plane.

Often one wants to "condition" interferometer data by shifting delay functions (e. g., to correct for clock errors), shifting frequencies or velocities (e. g., to correct for time varying Doppler shifts), or to change spectral resolution or numbers of spectral channels. As for the detection or fringe fitting problem, each conditioning operation is usually most easily accomplished in one of the four visibility planes.

For example, shifting in delay is often necessary to correct for clock errors. Usually the desired shifts are not integral multiples of the delay channel spacings, and, hence, the shifts require *interpolation* between channels. By use of the shift theorem for Fourier transforms, a shift in one domain results in a linear phase

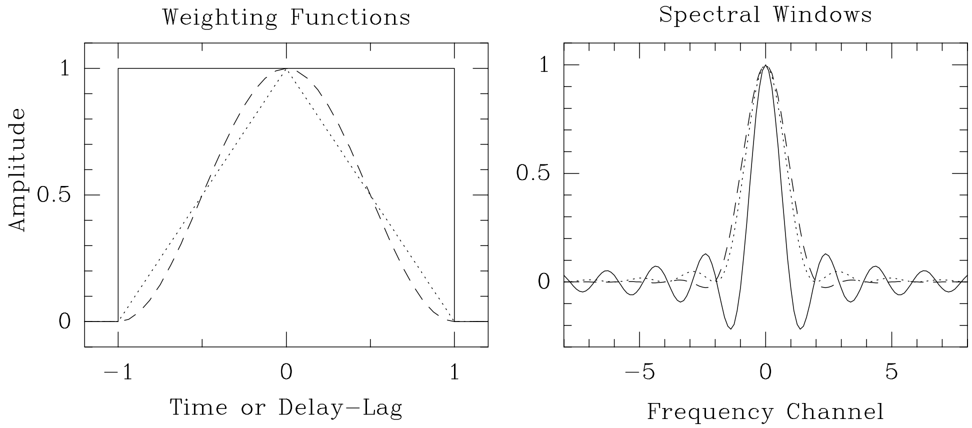


Figure 24–3. Weighting functions and spectral windows: The solid line depicts a uniformly (boxcar) weighted time or delay-lag function and its equivalent spectral response. The dashed line is for Hanning weighting. The dotted line gives the response for an FX correlator, whose *individual* telescope voltages are uniformly weighted. This process is equivalent to triangular weighting for an XF correlator.

slope in the other domain. Thus, delay shifts are most easily accomplished in the frequency domain as follows:

1. Fourier transform from the delay axis (in either plane "a" or "b") to a frequency axis (in plane "c" or "d").

2. Multiply the frequency functions (spectra) by $e^{i2\pi\nu_n\Delta\tau}$, where ν_n is the (video) frequency of the nth channel and $\Delta\tau$ is the desired delay shift.

3. If needed, Fourier transform back to the delay domain.

For a line source, the shape of the spectrum in the frequency domain can be arbitrarily complicated. Indeed, the spectrum shown in Figure 24–2 was generated by blending many narrow lines with different visibility phases, corresponding to different positions on the sky.

Now, we consider how a single, narrow (spectrally unresolved) line would appear in the correlator output spectrum. In an XF correlator, a *finite* delay-lag function is generated and later Fourier transformed into a spectrum. If a uniform weighting function is applied to the data before transforming, this is equivalent to multiplying the "complete," or infinitely long, delay-lag function by a rectangular function that is 0 outside the transform range and 1 inside that range (see the solid line in the left panel of Figure 24–3). Since the Fourier transform of a product of two functions is the convolution of the Fourier transform of each function, this procedure of multiplying the "complete" delay-lag function by a weighting function, results in convolving the true source spectrum with a "window" function. For uniform weights, the window function is the Fourier transform of the rectangular weighting function and is a $\sin(x)/x$ (see the solid line in the right panel of Figure 24–3). This leads to both positive and negative "side-lobes" in the frequency channel domain and is the correlator response for a narrow spectral line. One alternative weighting function, called Hanning weighting, is given by the dashed lines in Figure 24–3. By down-weighing the larger delay-lags, one can reduce side-lobe levels in the frequency domain at the expense of degrading the resolution of the main peak.

For the case of the FX correlator, a time sequence of voltages for a given telescope is Fourier transformed and then spectra from a pair of telescopes are cross multiplied. If the time domain data are uniformly weighted, the spectral window for each telescope's spectrum is a $\sin(x)/x$ function. Cross-multiplying the spectra from two telescopes yields a $\sin^2(x)/x^2$ function, which, as shown by Moran (1976), is equivalent in the time domain to triangular weighting (dotted lines in Figure 24–3). Note the similarity of the Hanning window for the XF and the "uniform" weighted window for the FX correlators.

3. Spectral-line Calibration

I now will outline how a typical spectral-line VLBI observation is conducted and analyzed. This hypothetical observing program will follow a repeated sequence in which a compact continuum source, whose position is accurately known, is observed for a brief period of time, followed by a spectral line source, which is usually observed for a longer duration. Continuum sources, or calibrators, will be used to determine the behavior of the station clocks, and they can also be used to solve for baselines and atmospheric parameters. In addition, the continuum sources can be used to obtain "off-source" time for total-power (auto-correlation) analysis. Since the continuum sources usually have no detectable spectral lines at the maser frequencies, the auto-correlation spectrum on such a source yields the amplitude response of the bandpass and allows one to generate "on–off/off" total-power spectra for calibration purposes.

For simplicity, and because most VLBI data has been processed with XF correlators, I will assume that the output of the correlator will be a set of cross-correlation and auto-correlation functions for each correlator integration period, Δt, of typically ~ 1 sec. Thus, the data start as represented by plane "a" of Figure 24–2. Data analysis proceeds as described below.

3.1. Instrumental Parameters

Here we find various instrumental parameters responsible for phase shifts in the complex fringe visibilities. If these phase shifts are not removed from the data they will degrade the quality of the line (channel) maps and any astrometric measurements. Commonly the instrumental parameters involve the time variation of the clock at each telescope (often modeled as a linear or slowly varying function of time), the electronic phase differences among video bands, and the phase variations across each video band. In some cases, station coordinates (baselines) and atmospheric models are corrected based upon the calibrator observations.

Most of the necessary calibration information can be obtained by fitting fringes to the continuum data. This is most easily accomplished by Fourier transforming the time axis of plane "a" to the fringe-rate axis of plane "b", concentrating all the "fringe power" contained in the time spanned by the Fourier transform in a few fringe-rate pixels. Next one finds, via a suitable fitting algo-

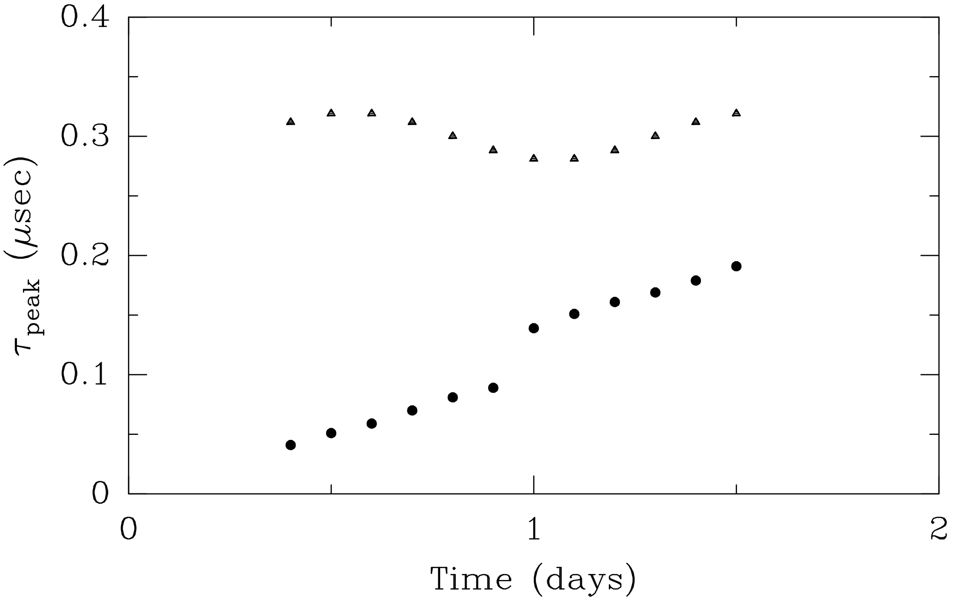


Figure 24–4. Interferometer delay versus time. Filled circles show typical behavior reflecting a relative clock drift of $\approx 0.1\mu$sec/day and a clock jump (discontinuity), possibly caused by a station power failure. Filled triangles show a sinusoidal variation with a 24 hour period, indicating a significant parameter error in the correlator model.

rithm, the peak in the delay–fringe rate plane. With perfect clocks, baselines, and atmospheric model parameters, the VLBI processor would exactly compensate for the delay and fringe-rate of the interferometer, and the fringe power would peak in the zero-offset channels of the delay–fringe-rate plane. If the fringes do not peak there, the residual interferometer delays are not zero and post-correlation corrections may be necessary. Shifts in the station clocks will manifest themselves as a shift in the location of the peak in the delay function. Clock drifts, baseline errors, and atmospheric variations result in both delay and fringe-rate offsets. Figure 24–4 provides two examples of the time variation of the residual interferometer delay, both requiring correction. Finally, electronic phase-shifts among bands, caused largely by mismatched low-pass filters and IF mixers, can be determined directly by comparison of the peaked fringe *phase* measured simultaneously among bands on a strong calibrator.

3.2. Doppler Tracking

We now shift attention to the data from the spectral line source. Soon we will want to Fourier transform these data from the delay domain of plane "a" to the frequency domain of plane "c". However, there is a correction that is commonly needed and which is most conveniently done in plane "a". During an observation the frequency (or Doppler velocity) of a given spectral line will change, owing to the commonly used procedure of using fixed local oscillators and the presence of frequency changes in the signals caused by rotation of the

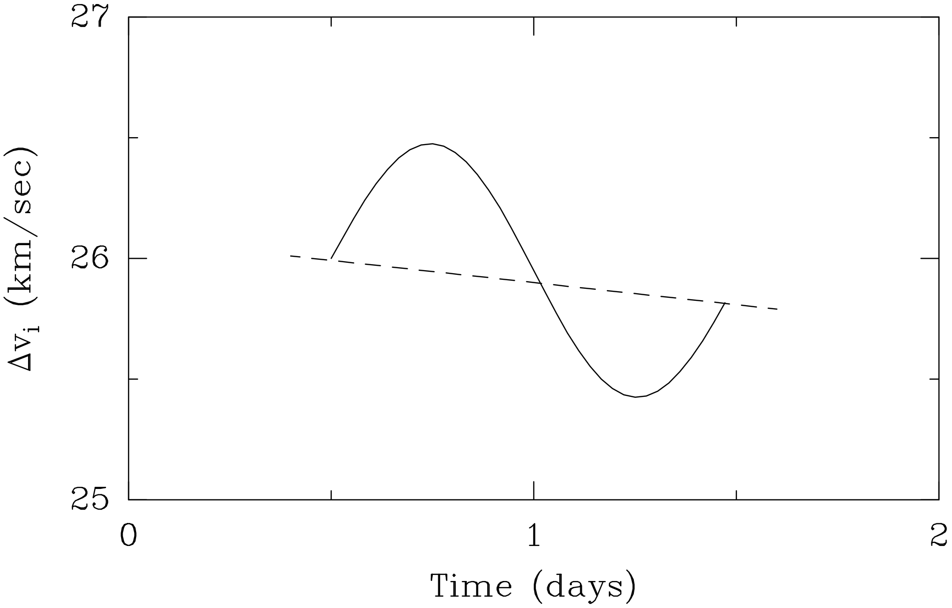


Figure 24–5. Doppler velocity of a radio source caused by the rotation of the Earth (solid line) and the Earth's orbit around the Sun (dashed line). The Earth's orbital velocity, projected along the line-of-sight to the radio source, can approach 30 km/sec; while the Earth's rotation, similarly projected, can approach 0.5 km/sec.

Earth and its orbit about the Sun (e. g., Figure 24–5). If the correlator does not correct for these effects (so-called Doppler tracking), it is important to do so in the post-correlation processing. Otherwise a spectral line will slowly drift across the spectrum, degrading the spectral resolution in a systematic, time-dependent, fashion.

One can frequency-shift data by multiplying the delay functions by $e^{i2\pi\Delta\nu\tau_\ell}$, where $\Delta\nu$ is the desired frequency shift and τ_ℓ is the delay value:

$$V(t_k,\tau_\ell) \leftarrow V(t_k,\tau_\ell)\ e^{+i2\pi\Delta\nu\tau_\ell} \quad . \qquad (24\text{–}4)$$

In this and subsequent relations, the left arrow (←) is equivalent to an equals sign (=) in C or FORTRAN and indicates that the value of the right-hand side replaces that of the left-hand side. If $\Delta\nu$ as a function of time is calculated correctly, this correction will hold the spectral line steady in a given spectral channel for the entire observation period, once the data are Fourier transformed from the delay to the frequency domain (plane "c").

3.3. Transforming from Delays to Frequencies

Once all necessary calibration information has been obtained from the delay data, and Doppler tracking corrections have been made to the line source data, it is time to Fourier transform the data set to the frequency–time domain. The complex fringe visibilities in the delay domain, $V(t_k,\tau_\ell)$, are transformed to spectra, $S(t_k, \nu_n)$, by a discrete Fourier transform:

$$S(t_k, \nu_n) = \frac{1}{\mathrm{M}} \sum_{\tau_\ell = -\mathrm{M}\Delta\tau/2}^{+\mathrm{M}\Delta\tau/2} V(t_k, \tau_\ell)\; e^{+i2\pi\nu_n\tau_\ell}. \qquad (24\text{–}5)$$

Usually a Fast Fourier transform (FFT) algorithm is used to speed up this calculation. While previous VLBI correlators have usually produced delay functions, the VLBA correlator performs the Fourier transform on-line (before cross-correlating) and also is capable of Doppler tracking sources. Thus for the VLBA case, the steps described in sections 24.3.2 and 24.3.3 are handled in the correlator.

3.4. Spectral Line Amplitude Calibration

The interferometer spectra in plane "c" are in "correlation-coefficient" units: that is, they are an effective interferometer antenna temperature divided by a system temperature. We wish to convert these to flux density units (e. g., Janskys) by calibrating the gain of the antennas, the system temperature of the receivers, and the effects of atmospheric emission and attenuation. For most continuum applications this is done by multiplying the correlation data by an effective system temperature (corrected to above the Earth's atmosphere), $T^*_{\rm sys}$, and the antenna gain, Q (in units of Jy/K). For an unpolarized source, $Q = 2k/A_{\rm eff}$, where k is Boltzmann's constant and $A_{\rm eff}$ is the effective collecting area of the antenna.

In the absence of digitization and sampling losses in the correlator, the product $T^*_{\rm sys}Q$, the system equivalent flux density (SEFD) in Janskys, scales correlation coefficients to fringe amplitudes in Janskys. For interferometers, as opposed to single antennas, the scaling is done by multiplying the correlation coefficients by the *geometric mean* of the SEFDs of the two antennas that were cross correlated. Explicitly,

$$S_{ij}(Jy) = b\,\sqrt{(T^*_{\rm sys}Q)_i\;(T^*_{\rm sys}Q)_j}\; S_{ij}(correlator), \qquad (24\text{–}6)$$

where b compensates for digitization and sampling losses, and i and j indicate the cross-correlation of the signals from the i^{th} and j^{th} antennas, respectively. For most continuum VLBI applications, the effective system temperatures ($T^*_{\rm sys}$) are measured by comparison of the total-power level on-source when a noise (cal) source of known temperature (or power, kT) is switched on and off. Similarly, the gain (Q) of each antenna is measured by comparing the total power level on- and off-source (using the program source if it is strong enough), also with the aid of the noise source.

For spectral line applications, it is often possible to measure directly the SEFD in Janskys ($T^*_{\rm sys}Q$) from the *auto*-correlation spectra *on* and *off* the line source. First, one generates a total-power spectrum, $S(\nu_n)$, in the standard manner used for single-antenna spectral observations:

$$S(\nu_n) = \frac{S^{on}(\nu_n) - S^{off}(\nu_n)}{S^{off}(\nu_n)}, \qquad (24\text{–}7)$$

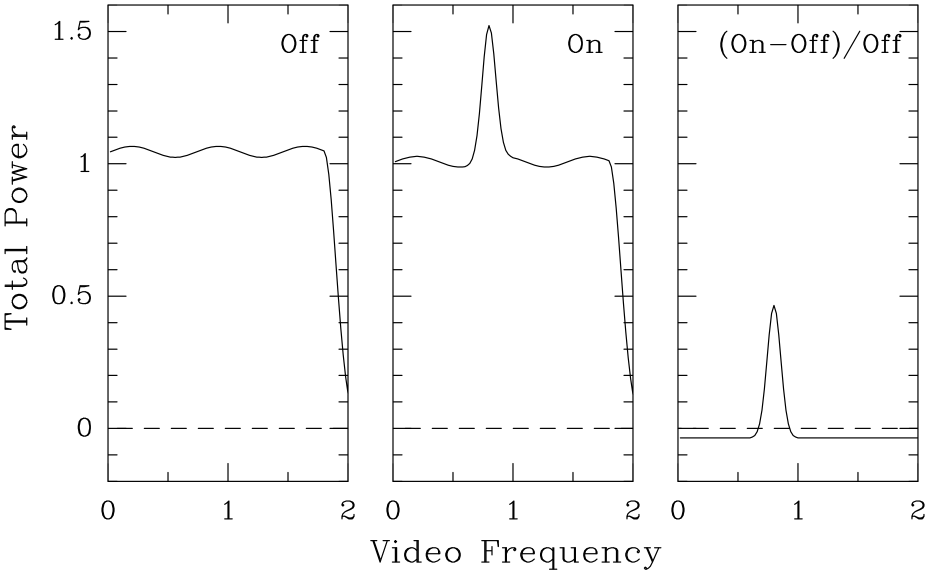


Figure 24–6. Total-power spectra from clipped (1-bit sampled) data. Shown is the power versus video frequency for a single telescope pointed off (left panel) and on (middle panel) a maser source. The off-source spectrum displays typical filter response including baseline ripple and high frequency roll-off. The normalized, total-power spectrum, formed by differencing the on- and off-source spectra and dividing by the off-source spectrum (right panel), is corrected for filter response and is proportional to the true source spectrum. Both the off- and on-source spectra have "unity" average value, since they are derived from clipped signals whose zero-delay auto-correlation value is unity, and hence the "on-off/off" spectrum has zero area (and a baseline level below zero).

where $S^{on}(\nu_n)$ and $S^{off}(\nu_n)$ are the raw on– and off–source auto-correlation spectra, respectively. This total-power spectrum will have "correlation coefficient" units and will be sensitive to telescope gain, system temperature, and atmospheric emission and absorption. For example, a spectral line in the total-power spectrum will have a lower strength if the antenna gain decreases, perhaps owing to a pointing error, or if the system temperature increases because a cloud passes over the antenna. One can quantify this performance change by comparing a total-power spectrum from each antenna for each observing scan against a single "template" total-power spectrum of the source. Often the template spectrum is obtained from the largest and best calibrated antenna used in the VLBI array. Each total-power spectrum is stretched (or shrunk) in amplitude until it best matches the template spectrum. The stretching (or shrinking) factor is directly proportional to the desired SEFD ($T^*_{\rm sys}Q$). Indeed, if the template spectrum is properly calibrated in Janskys, the stretching factor is precisely $T^*_{\rm sys}Q$.

Obtaining the gain calibration information for each antenna from its auto-correlation data has been widely used for spectral-line observations. It does require that some off-source observing time be scheduled into the program (and the data recorded on the VLBI tapes). Often one can use the observation time on continuum calibration sources for off-source data, since these data are only used to determine the shape of the passband. Fitting auto-correlation spectra can be an extremely precise method of achieving the *relative* calibration of fringe amplitudes among the antennas of a VLBI array. (*Absolute* calibration depends on the accuracy of the flux density scale assigned to the template spectrum.) Of course, it is important that sufficient signal-to-noise be achieved in the auto-correlation spectra. For a strong maser source precisions of better than 1% are common.

Potential problems with this technique are those that affect most single-dish spectroscopy and include "birdies" (interference) and spectral baseline ripples. Indeed, if auto-correlation data from the continuum calibration sources are used to construct "off-source" spectra, baselines may be poor as the on- and off-source spectra may not be acquired in angular or temporal proximity. Also, there is one subtle electrical problem than is not detected with auto-correlation. If the local oscillator used at the VLBI station does not have a pure "delta function" signal, but has power elsewhere (e. g., in 60 Hz harmonics), then the cross-correlated signal will be reduced (e. g., 60 Hz is outside the "fringe window" of data averaged to ≈ 1 second in the correlator). However, the auto-correlation signal will not be reduced, since all the power will usually be contained within one spectral channel (typically many kHz wide).

The auto-correlation data can be useful for more than calibration, since they indicate the total-power emitted by the source (at least all the power within the single antenna beam, which usually is arc-minutes in size). Interferometric data generally do not sample short baseline spacings. For connected-element interferometers like the VLA this occurs because antennas cannot get closer together than the distance at which they touch. For VLBI data the shortest spacings by definition are large (e. g., hundreds of km). One can, in principle, add in the auto-correlation data at the center of the (u, v)-plane (i. e."zero spacing" data) in the mapping step. However, in practice this is useful only if the fringe amplitudes on shorter VLBI spacing have a reasonable fraction of the total-power flux density.

3.5. Clock and Coordinate Corrections

Now the data are in the frequency–time domain (plane "c" of Figs. 1 and 2); they have been *amplitude* calibrated and are in units of Janskys. It is time to correct the visibility *phases* for various instrumental and propagation effects. Any errors in the station clocks, as well as possible coordinate errors in the locations of the antennas, will lead to time-dependent delay shifts. A delay shift appears as a translation along the τ-axis in data planes "a" and "b". After Fourier transforming the delay (τ) to the frequency axis (ν), *a delay shift appears as a linear frequency-dependent phase change.* Figure 24–7 plots two delay-functions, one whose peak is shifted from the center channel, and the resultant phase responses in the frequency domain.

From the calibration information obtained in section 24.3.1, we can generate a table of (or formulae for) the desired delay corrections to be applied to each interferometer baseline at any observing time. The complex fringe visibility spectrum, $S(t_k, \nu_n)$, can be corrected for a delay shift, $\Delta\tau$, as follows:

$$S(t_k, \nu_n) \leftarrow S(t_k, \nu_n)\ e^{-i2\pi\Delta\tau\nu_n}. \tag{24–8}$$

This process removes a phase slope across the video band caused by the delay error.

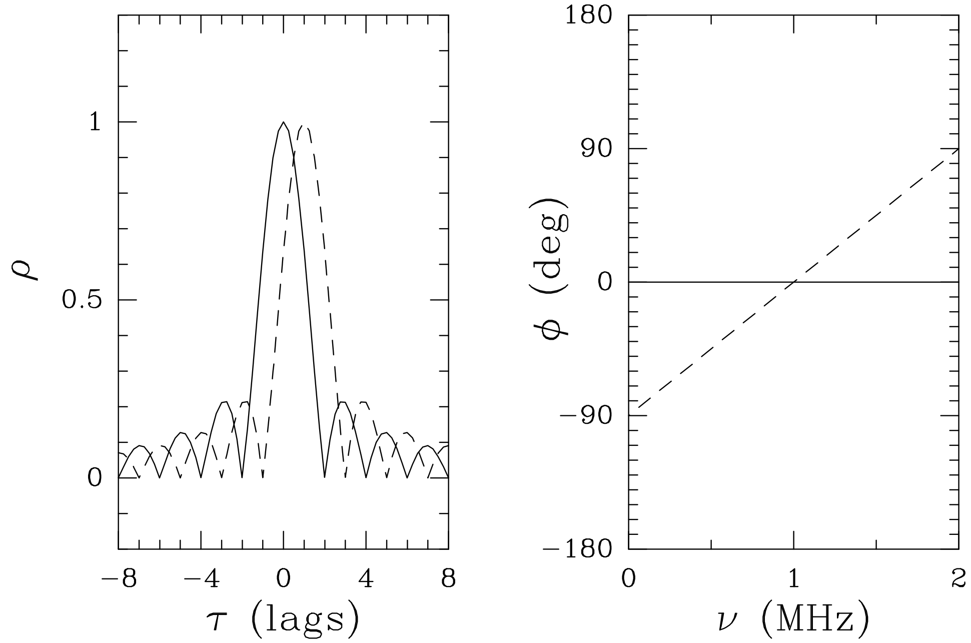


Figure 24–7. Interferometer correlation functions (left panel) and associated phase response versus video frequency (right panel). Shown are two cases in which the residual (post correlation) delay has (1) zero offset (solid line) and (2) a 1μsec offset. A 2 MHz video bandwidth and 1/4 μsec delay steps are assumed. Note that a residual delay error of one delay step results in a phase slope of 180 degrees across one video sideband.

If delay corrections involve only a correction for a *constant* clock offset at each station, then one can skip to the next step in the calibrations process. However, if, for example, the station clocks drift in time (caused by a frequency error in the oscillator driving the clock), or coordinate corrections are needed, then one must also correct the visibility phase by a video-frequency *independent* amount as follows:

$$S(t_k, \nu_n) \leftarrow S(t_k, \nu_n)\ e^{-i2\pi\Delta\tau\nu_0}, \qquad (24\text{–}9)$$

where ν_0 is the local oscillator frequency, or sky frequency of the zero video-frequency channel, of the observing band.

One other calibration step is associated with clock or coordinate corrections. When any time-dependent phase correction is made, the original data probably had a non-zero residual fringe-rate. Such data have been degraded in amplitude because they were time averaged (in the correlator) over a portion (hopefully small) of the fringe-rate cycle. Averaging of the fringe visibility, $S(t, \nu)$, which varies in time sinusoidally at a fringe-rate, f, decreases its amplitude by a factor given by $\sin(\pi f T)/(\pi f T)$, where T is the integration or averaging time. Thus, when one makes time-dependent phase corrections (equivalently fringe-rate corrections), one should also consider correcting fringe amplitudes by the reciprocal of this "$\sin(x)/x$ " function. Of course, such a correction can only be done correctly for one position on the sky (i. e., the phase center of the map).

3.6. Electronic Phase Shifts

Many spectral line and continuum VLBI observations use a number of narrow bands recorded separately on tape to span a wider bandwidth on the sky. Combining the fringe visibilities from different bands requires correcting the data for the visibility phase differences among the bands. Such phase differences are primarily caused by the use of different IF mixers and different electronic components in the low-pass (video) filters following the final mixing stage. The calibration data compiled in section 24.3.1 is now applied to the bands to align their phases to that of a single (reference) video band. This is accomplished as follows:

$$S(t_k, \nu_n) \leftarrow S(t_k, \nu_n)\ e^{-i(\phi^N - \phi^R)}, \tag{24–10}$$

where $(\phi^N - \phi^R)$ is the *band-averaged* phase difference between video band "N" and the reference video band "R".

Now that the *band-averaged* phases for all video bands are aligned, one can consider correcting for the (typically $\sim 10°$) phase ripples across each individual band. This is usually done only for very precise astrometric measurements where very high dynamic range is needed in the channel maps. The phase ripple in each band can be seen on a single scan on a very strong continuum source or bandpass calibrator. Such a correction is accomplished, mathematically, as follows:

$$S(t_k, \nu_n) \leftarrow S(t_k, \nu_n)\ e^{-i\phi^N(\nu)}, \tag{24–11}$$

where $\phi^N(\nu)$ represents the phase as a function of frequency across video band "N".

3.7. Phase Referencing

At this point the data are in the form of amplitude calibrated spectra (in plane "c"), and they have been corrected for clock and coordinate errors and electronic phase shifts. Were this VLA data, taken in a small configuration under good weather conditions, all that would remain to do would be to apply a slowly varying phase correction, determined from continuum calibrator scans, to remove the effects of visibility phase fluctuations caused by the *difference* in the propagation of the radio waves through the Earth's atmosphere for the different antennas. However, VLBI data usually contains very large atmospheric phase fluctuations (since the atmospheric fluctuations are largely uncorrelated among the antennas) and an additional term owing to the use of *independent* local oscillators for each antenna. Therefore, most VLBI data must be "phase referenced" in some manner.

For continuum VLBI observations the usual procedure is to attempt "self-calibration," although rapid switching between the program and calibration source has recently shown great promise for calibrating the interferometric phase for the VLBA. Self-calibration uses the signals from the program source itself, coupled with some knowledge of the source structure, to iteratively solve for the atmospheric or local oscillator phase fluctuations as well as the source structure. However, for spectral-line VLBI and, in particular, for observations of cosmic masers, one can often use the phase variation from a spectral channel containing simple source structure as a reference. Ideally one desires a strong, unblended spectral component as a reference. This may not be available, but the steep edge

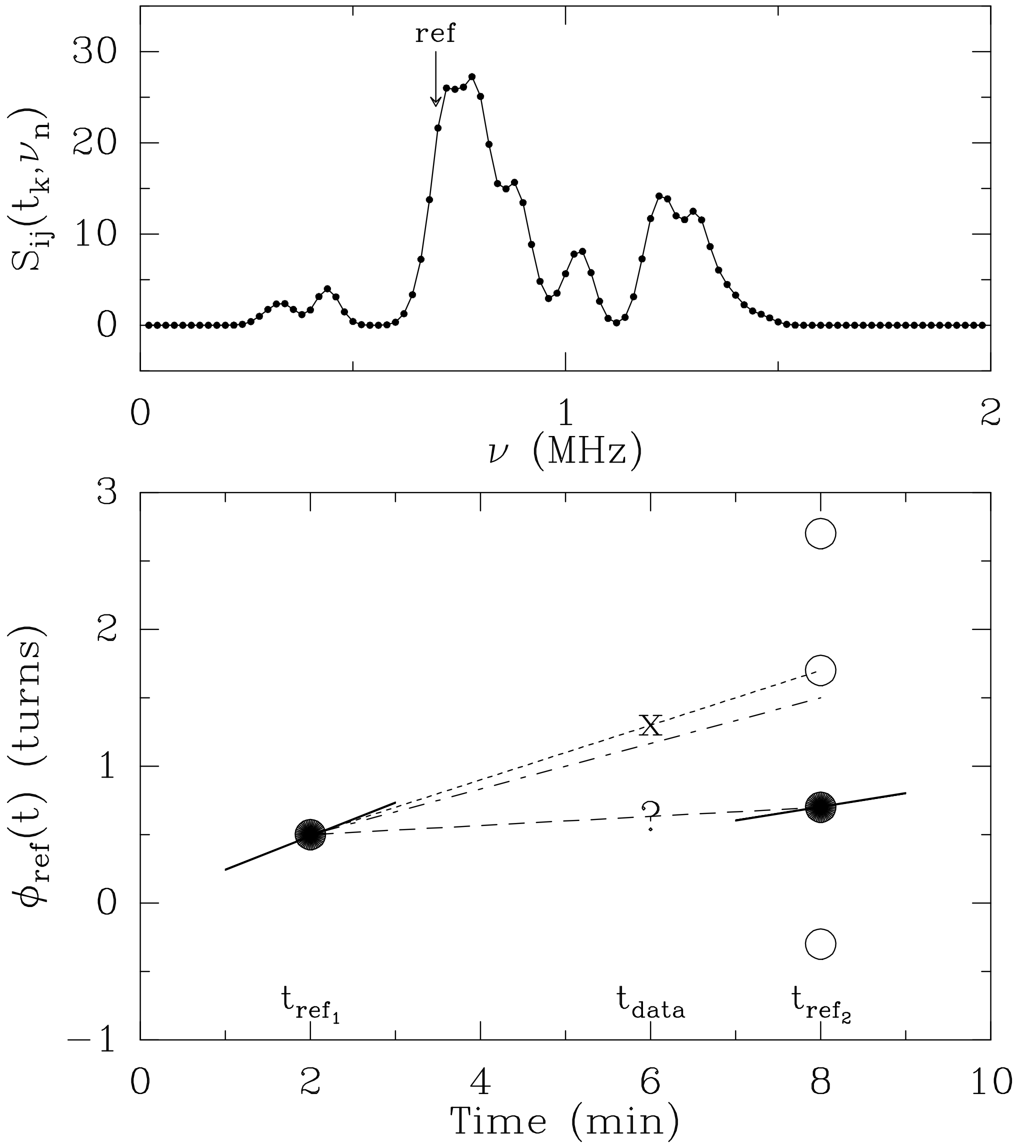


Figure 24–8. Phase-referencing. *Top:* Interferometer spectrum indicating a ("ref") channel to be used as a phase reference. *Bottom:* Interferometer phase vs time. Solid circles indicate the phase, arbitrarily defined between 0 and1 turn, derived by fringe fitting data from the "ref" channel at times t_{ref_1} and t_{ref_2}. Heavy lines through the solid circles indicate the fitted fringe rates. The dashed line is a direct connection of the two fitted phases; "?" denotes phase interpolated via this connection at time t_{data}. This phase connection and the interpolated phase are incorrect. Phases have integral-turn ambiguities and could be as indicated by the open circles. Averaging the fringe rates from the two fits suggests that the phase should be extrapolated along the dash-dotted line and that, relative to the phase at t_{ref_1}, the phase at t_{ref_2} should be one turn higher (i. e., 1.7 turns). This resolves the turn ambiguity between the two times. The best interpolation should follow the dotted line, yielding the phase indicated by the "x" for data at time t_{data}.

of a blend of lines (as in Figure 24–8) often provides an adequately clean signal. In any event, one can "self-calibrate" the reference channel data, if needed, to correct for complex structure in that spectral channel.

For simplicity, assume that spectral channel "r", at frequency ν_r, contains a strong signal from a single point source on the sky. One can copy the visibility spectra, $S(t_k, \nu_r)$, from this channel in plane "c", into a temporary data base, and Fourier transform (the time domain into the fringe-rate domain) to plane "d". The next step is to fringe-fit this data over a time period (typically minutes) long enough to provide an adequate signal-to-noise (SNR) ratio ($>$ 3 to 1) to measure phase precisely, but short enough so as not to exceed the coherence time of the interferometer. (Note that the phase error is $\approx 1/\mathrm{SNR}$ radians.) Should one choose either too short or too long a time period for fringe fitting the reference channel data, noise will be introduced in the referencing procedure and lead to degraded maps. In some cases, the time required to achieve an adequate SNR exceeds the coherence time of the interferometer and phase referencing is not possible!

The procedure described above yields a table of fringe-rates and phases as a function of time for all interferometer baselines. Almost all phase fluctuations in this table can be thought of as caused by something happening at each antenna (i. e., antenna based), with little or no additional contribution that is specifically baseline, or correlator, dependent. In this case, a "reference" antenna can be chosen (somewhat arbitrarily) and individual antenna-based tables of phase fluctuations determined. These antenna-based tables will have the property that differencing the phases from antennas i and j will yield the phases observed on the (i, j)–baseline. In practice, one can directly solve for station-based reference phases via "global fringe fitting" techniques.

The final step in the calibration process involves using these antenna-based tables of visibility phase as a function of time to predict, and remove, the phase fluctuations on all interferometer baselines. The phase from the calibration table for antenna i is interpolated to the time, t_{data}, of a visibility record in the full data base, yielding $\theta_i(t_{\mathrm{data}})$. When performing this interpolation it is important to resolve integral turn ambiguities between entries in the reference feature phase table (see Figure 24–8). One can then phase reference all the data for the baseline from the i^{th} to the j^{th} antenna, $S_{ij}(t_k, \nu_n)$, as follows:

$$S_{ij}(t_k, \nu_n) \leftarrow S_{ij}(t_k, \nu_n)\ e^{-i(\theta_i - \theta_j)}, \qquad (24\text{–}12)$$

At this point we have completed the calibration of the spectral line data and mapping can proceed.

4. Fringe Rate Mapping

Once the data are calibrated, one can often proceed to construct line-channel maps in a manner similar to that for continuum sources. However, what if one has very limited data? For example, data from a few baselines or data with poor amplitude calibration are difficult to analyze via standard synthesis mapping techniques. In addition, what if one has emission spread over a very large (and possibly unknown) field on the sky? For example, the H_2O masers in the Orion-KL region are found over a 40 × 40 arcsec field. Using an interferometer with

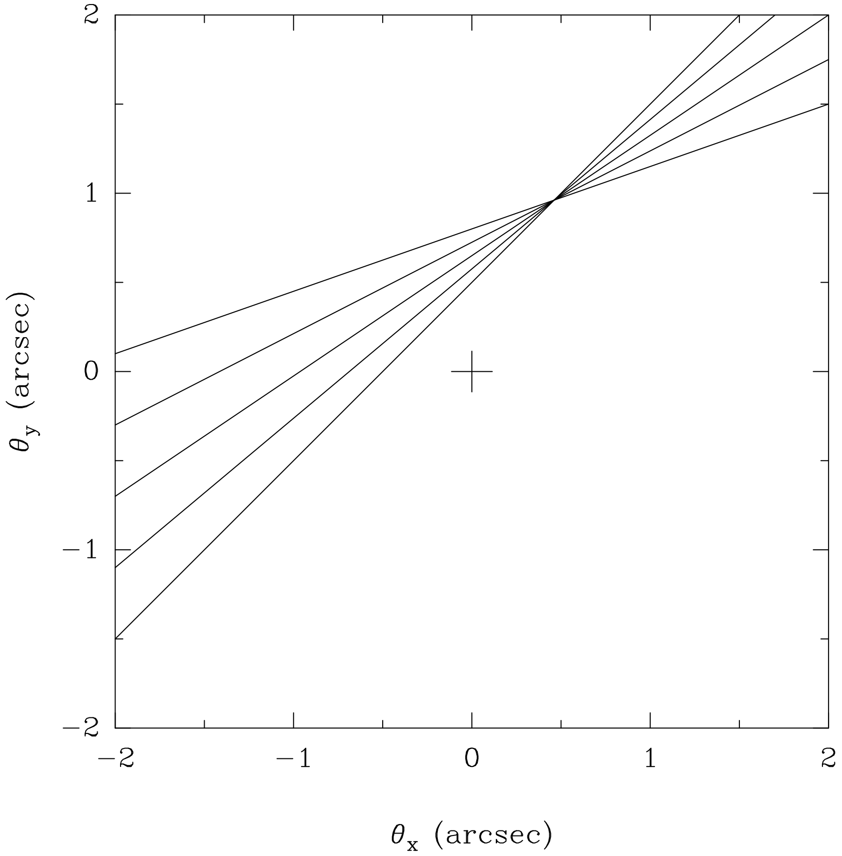


Figure 24–9. Fringe-rate mapping. Shown is a 4 arcsec square field of the sky. The origin of the field, indicated by the cross, is the location of the emission from the channel used to phase reference the data. Each line on the figure gives positions on the sky consistent with a fringe-rate measurement in a single spectral channel at a given time on one interferometer baseline. Measurements at different times or on multiple baselines result in lines of different orientation, and the intersections of the lines locates the position of the emission.

a synthesized beam of 1 mas, and mapping with 3 pixels per beam, would require maps with $120,000 \times 120,000$ pixels! Currently this is computationally "challenging."

For the reasons outlined above, one often wants a "first cut" or crude mapping technique that is not very sensitive to (u,v)–coverage or amplitude calibration. One such technique is called fringe-rate mapping. Fringe-rate mapping makes use of the fact that a single fringe-rate measurement (determined by fringe fitting data in a spectral channel at a given time and on a given baseline) constrains the location of the dominant emission to a "line" on the sky. Figure 24–9 gives an example of 5 such fringe-rate lines; these lines might come from 5 scans on one baseline or from 1 scan on 5 baselines. The intersection of the fringe-rate lines indicates the location of the emission (relative to the position of the emission used to phase reference the data).

With fringe-rate mapping, one can map large fields with a modest computational resources. Amplitude calibration is not necessary, since such calibration does not strongly affect fringe rate estimates for most fringe fitting algorithms. Typical accuracies of fringe-rate mapping for strong H_2O masers are ~ 1 mas. Note, however, that as for synthesis mapping, limited time or baseline coverage often leads to limited positional accuracy. In the example shown in Figure 24–9, positional accuracy is better perpendicular to the fringe-rate lines (from the upper-left to lower-right) and poorer along the fringe-rate lines.

References

Moran, J. M. 1976, in *Methods of Experimental Physics*, Ed. M. L. Meeks, (Academic Press; New York), 174–197.

Thompson, A. R., Moran, J. M. & Swenson, G. W. 1986, *Interferometry and Synthesis in Radio Astronomy*, (John Wiley & Sons; New York)

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

25. VLBI polarimetry

A. J. Kemball
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. This lecture describes the calibration and imaging of polarization VLBI data, with special reference to those aspects of calibration which differ from standard techniques employed in conventional total intensity VLBI observations. Both spectral-line and continuum observations are considered, and brief mention is made of more specialized cases, including polarization VLBI using orbiting antennas. A summary is provided of general recommended practices when scheduling polarization VLBI observations, and a discussion given of the sources of calibration error which may limit the fidelity of the final polarization images.

1. Introduction

Polarimetry is an integral component of radio interferometry; the same basic principles of interferometry apply when imaging the incident radiation in full polarization, as characterized by the four Stokes parameters $[\mathcal{I}, \mathcal{Q}, \mathcal{U}, \mathcal{V}]$, as do when imaging in total intensity alone. Polarization VLBI is no different in this respect, and constitutes a similar generalization of the basic principles that apply to VLBI observations sensitive only to a single polarization, albeit with increased complexity in instrumental calibration.

By allowing radio-astronomical emission to be imaged in all Stokes parameters at millliarcsecond (mas) angular resolution, polarization VLBI is a powerful tool, adding important additional information for the astrophysical interpretation of the underlying physical processes in compact sources. Such observations can be carried out for both spectral-line and continuum sources, with the subset of relevant Stokes parameters dependent on the underlying astrophysical questions being investigated. In continuum observations of active galactic nuclei, for example, circular polarization (Stokes V) may frequently be ignored, but it is highly relevant in polarization VLBI observations of certain astrophysical masers.

The fundamental scientific contribution of polarization VLBI is the opportunity it offers to study the magnetic field distribution in radio sources at mas resolution. It also adds information on the underlying emission mechanisms in such sources, as well as the radio-frequency propagation properties of the medium in the immediate source environment and along the line of sight to the observer. There are also more specialized uses of polarization VLBI which will be covered later in this lecture.

For a discussion of the general principles and underlying theory of interferometric polarimetry the reader is referred to other lectures in this school, specifically Lectures 6 and 32, as well as other references including Thompson, Moran & Swenson (1986) and Fomalont & Wright (1974). The technique was demonstrated first for connected-element arrays (Morris, Radhakrishnan & Seielstad 1964, Conway & Kronberg 1969, Weiler 1973). Calibration of polarization VLBI data was reported first by Cotton et al.(1984) and Roberts et al.(1984). Subsequent work describing technical developments in polarization VLBI can be found in a series of papers by Cotton (1989), Roberts, Brown & Wardle (1991), Cotton (1993), Roberts, Wardle & Brown (1994), Kemball, Diamond & Cotton (1995), and Leppänen, Zensus & Diamond (1995).

Polarization VLBI studies of compact extra-galactic radio sources have provided numerous insights into the magnetic field morphology and physical conditions in the inner regions of active galactic nuclei (Wardle & Roberts 1994; and references therein). Polarization VLBI reveals differences in source properties that may not be discernible in observations that are not sensitive to linear polarization. In particular, centimeter polarization VLBI surveys of compact extragalactic sources have revealed clear morphological differences between BL Lac objects and quasars, when considered as separate optical classes, in respect of the orientation of the electric vector position angle with respect to the underlying radio jet (Gabuzda et al.1992, Cawthorne et al.1993). Polarization VLBI has also allowed the investigation of Faraday rotation on parsec scales in active galactic nuclei (Taylor 1998), thus providing important information on conditions in the narrow line regions of such objects.

Polarization VLBI observations of spectral line sources have primarily concerned astrophysical masers, as found in HII regions and in the circumstellar environments of late-type stars. Polarization VLBI results for such sources have been reported by Garcia-Barreto et al.(1988), Bloemhof et al.(1992), Kemball & Diamond (1993) and Kemball & Diamond (1997), amongst others. These observations allow the determination of the strength and morphology of the magnetic field in the masing regions, and thus provide important information on the associated dynamics. In addition, such observations provide insight into the emission and transport mechanisms for polarized maser radiation, an active area of current theoretical research. Polarization information of mas scales is also helpful in identifying matching maser components in proper motion studies over multiple epochs.

The objective of this lecture is to cover those aspects of calibration that are unique to polarization VLBI, and to address issues relevant to scheduling and reducing such observations in practice. Improvements in the polarization performance and sensitivity of VLBI networks in recent years have rendered this technique accessible to all VLBI observers, and this lecture aims to stress this point. This is especially true for polarization observations using the VLBA network, which has homogeneous polarization performance across the network as a whole.

The basic questions concerning the calibration and imaging of polarization VLBI data are covered in Sections 2 and 3, with a treatment of more specialized cases given in Section 4. A discussion of the practical aspects of polarization VLBI observing in provided in Section 5.

2. Polarization VLBI calibration

This section contains a description of the key calibration issues affecting polarization VLBI observations. This discussion is confined to those calibration matters which are unique to polarization VLBI; the calibration of general VLBI observations is discussed in Lectures 22 and 24.

2.1. Instrumental and propagation effects

Polarization VLBI is based on principles common to interferometric polarimetry in general. During observing, dual orthogonal polarizations are recorded at each antenna and all polarization cross-products are formed during subsequent correlation. The discussions that follows is confined to the case of dual circular polarization recording, with polarization products in this case of $[RR, LL, RL, LR]$. The Stokes parameters of the incident radiation above the atmosphere, are defined by correlations between the electric field in the usual way, here represented in a circularly-polarized basis $(\mathbf{E}^R, \mathbf{E}^L)$.

Propagation through the atmosphere, feeds and receiving electronics at each antenna introduces phase and amplitude offsets in each nominal recorded polarization, which in general need to be considered separately. The full model for all instrumental effects, including those introduced during correlation, is complex and dependent on the type of antenna and receiving electronics at each site. A representation for a generic interferometer, which is independent of instrumental configuration, is provided by the measurement equation formalism (Hamaker, Bregman & Sault 1996, Lecture 32). A condensed parametrization is adopted here, using the terms $G_m^{p,q}$, $B_m^{p,q}$, $D_m^{p,q}$ and γ_m^{p-q}, where the subscript m denotes the antenna number, and the superscript $[p, q] \in [RCP, LCP]$ indicates the nominal recorded polarization. The term $G_m^{p,q}$ defines the composite sum of all instrumental and atmospheric propagation effects remaining after correlation for the nominal recorded polarization p at antenna m, while $B_m^{p,q}$ defines the normalized complex bandpass response. The complex polarization leakage term D_m^p reflects the degree of instrumental contamination for polarization p from the orthogonal polarization q (Conway & Kronberg 1969), and is equivalent to an ellipticity-orientation representation of the feed response. The term γ_m^{p-q} is used here to denote the excess differential phase due to Faraday rotation in the ionosphere.

The measured correlations for each polarization pair in this formalism take the form (Kemball, Diamond & Cotton 1995),

$$
\begin{aligned}
r_{mn}^{RR}(u,v) &= (G_m^R G_n^{R*})(B_m^R B_n^{R*})[(\mathcal{RR})e^{-j(\alpha_m-\alpha_n)} \\
&+ (\mathcal{RL})D_n^{R*}e^{-j(\alpha_m+\alpha_n)}e^{j\gamma_n^{R-L}} + (\mathcal{LR})D_m^R e^{j(\alpha_m+\alpha_n)}e^{j\gamma_m^{L-R}} \\
&+ (\mathcal{LL})D_m^R D_n^{R*}e^{j(\alpha_m-\alpha_n)}e^{j(\gamma_m^{L-R}-\gamma_n^{L-R})}] \\
r_{mn}^{LL}(u,v) &= (G_m^L G_n^{L*})(B_m^L B_n^{L*})[(\mathcal{LL})e^{j(\alpha_m-\alpha_n)} \\
&+ (\mathcal{LR})D_n^{L*}e^{j(\alpha_m+\alpha_n)}e^{j\gamma_n^{L-R}} + (\mathcal{RL})D_m^L e^{-j(\alpha_m+\alpha_n)}e^{j\gamma_m^{R-L}} \\
&+ (\mathcal{RR})D_m^L D_n^{L*}e^{-j(\alpha_m-\alpha_n)}e^{j(\gamma_m^{R-L}-\gamma_n^{R-L})}] \qquad (25\text{–}1) \\
r_{mn}^{RL}(u,v) &= (G_m^R G_n^{L*})(B_m^R B_n^{L*})[(\mathcal{RL})e^{-j(\alpha_m+\alpha_n)} \\
&+ (\mathcal{RR})D_n^{L*}e^{-j(\alpha_m-\alpha_n)}e^{j\gamma_n^{L-R}} + (\mathcal{LL})D_m^R e^{j(\alpha_m-\alpha_n)}e^{j\gamma_m^{L-R}} \\
&+ (\mathcal{LR})D_m^R D_n^{L*}e^{j(\alpha_m+\alpha_n)}e^{j(\gamma_m^{L-R}-\gamma_n^{R-L})}] \\
r_{mn}^{LR}(u,v) &= (G_m^L G_n^{R*})(B_m^L B_n^{R*})[(\mathcal{LR})e^{j(\alpha_m+\alpha_n)} \\
&+ (\mathcal{LL})D_n^{R*}e^{j(\alpha_m-\alpha_n)}e^{j\gamma_n^{R-L}} + (\mathcal{RR})D_m^L e^{-j(\alpha_m-\alpha_n)}e^{j\gamma_m^{R-L}}
\end{aligned}
$$

$$+ \quad (\mathcal{RL})D_m^L D_n^{R*} e^{-j(\alpha_m+\alpha_n)} e^{j(\gamma_m^{R-L}-\gamma_n^{L-R})}]$$

where α_m denotes the parallactic angle at antenna m.

This equation is set in the visibility plane, with the quantities on the left denoting the measured data, as labelled after correlation. The true polarization properties of the incident electric field $[\mathcal{RR}, \mathcal{LL}, \mathcal{RL}, \mathcal{LR}]$ appear in terms on the right-hand side of these equations, also represented in the visibility plane. Calibration of polarization VLBI data involves the estimation of the unknown terms $G_m^{p,q}$, $B_m^{p,q}$, $D_m^{p,q}$ and γ_m^{p-q} to allow the determination of the true polarization correlations of the field. These are then converted to the image-plane Stokes representation $[\mathcal{I}, \mathcal{Q}, \mathcal{U}, \mathcal{V}]$, through Fourier transformation.

It needs to be stressed that a full calibration model may need to be more complex than this, including a dependence on time, frequency and angular position for all calibration terms, and the correct separation of visibility plane and image plane effects. However, the model adopted here is sufficient to define the key elements of polarization VLBI calibration.

The system of equations above is coupled; in practice approximations are adopted to break the coupling and linearize the system to a more manageable form. The approximations typically rely on the fact that the polarization leakage terms D_m^p are usually small, $|D| \leq 0.1$, and that the cross-polarization correlations are, in most cases, a small fraction of the parallel-hand correlations. In this form, terms of order $O(D^2)$ or $O(D.(Q \pm jU))$ are treated as second-order, and ignored.

2.2. Calibration

In the linearized formulation, the antenna-based calibration terms G_m^p and B_m^p can be determined separately in each nominal polarization from the parallel-hand data alone, using the same techniques employed in standard VLBI data reduction. An important additional requirement imposed by polarization VLBI is that all phase and amplitude offsets be determined between the calibration terms in each recorded polarization. Alternatively stated, the phase of the G_m^p and B_m^p terms needs to be referred to the same reference polarization at the reference antenna, and the amplitude corrections in each polarization placed on the same absolute scale. The determination of these offsets is described below. In addition, the phase correction needs to be determined in such a way that positional coincidence is maintained on the sky for each polarization. As is well-known from conventional VLBI reduction, angular position offsets can be absorbed into phase correction terms during self-calibration. This is explained by the standard Fourier transform shift theorem. For recording in (RCP,LCP), this problem does not arise if the source is circularly unpolarized, thus allowing the use of a circularly unpolarized source model during self-calibration. This assumption is made for the remainder of this section. The opposite case is described in a later section dealing with polarization calibration of spectral-line VLBI data.

Once the terms G_m^p and B_m^p have been referred to the same calibration reference frame, they can be used to correct all polarization correlation pairs. In the linearized model, the calibrated cross-polarized data are then used to determine the polarization leakage terms D_m^p, as discussed in the section on feed calibration.

Parallactic angle. As shown in Equation 1, there is a parallactic angle phase contribution in the expressions for all polarization correlation pairs. The dominant contribution can be factorized as an antenna-based phase correction equal in magnitude to the parallactic angle but of opposite sign for each nominal circular polarization. If applied at the outset of calibration, this antenna-based correction removes the parallactic angle term for the primary source contribution in the expression for each polarization. This step is required to simplify subsequent phase and polarization calibration as discussed below.

Correlator corrections. For polarization VLBI, all corrections required to account for digital signal processing effects in the correlator need to be made as in the case for conventional VLBI, and they are not discussed in detail here. In general, correlation is a polarization independent process as the correlator model invariably does not contain polarization-dependent terms.

For multi-level quantization schemes of digital sampling (2-bit or more), amplitude offsets may arise between the recorded polarizations if they are systematically assigned different samplers. This may arise on the VLBA as polarizations are typically recorded in alternating baseband converters, which mirrors the assignment of samplers to odd and even converter units. Higher quantization samplers introduce amplitude offsets if the threshold voltage levels deviate from their nominal values. These offsets need to be corrected in any event, either with reference to mean autocorrelation levels or through the use of sampler state-count data. However, as pointed out here, these offsets can introduce polarization-dependent amplitude offsets, if uncorrected in certain cases and this correction is a recommended step in polarization VLBI data reduction.

Amplitude calibration. For polarization VLBI, initial amplitude calibration is carried out using the same techniques employed in standard VLBI data reduction. For continuum data, the amplitude correction factors are derived from the recorded T_{sys} or T_{ant} measurements in each polarization, and *a priori* knowledge of the point source sensitivity of the antenna in each polarization as a function of position on the sky. For spectral-line data, the amplitude correction factors may be derived from the parallel-hand autocorrelation data using the template spectrum method (Reid et al.1980). In this case, *a priori* gain and system temperature measurements are still required to set the absolute amplitude scale for the template spectrum in each polarization separately.

The initial *a priori* amplitude calibration is thus vulnerable to offsets between the two recorded polarizations as may be introduced by systematic errors in the measured calibration data. These offsets can be determined from careful amplitude self-calibration of a source with zero circular polarization, as is effectively the case for most continuum compact extra-galactic radio sources. In this case the ratio of the residual amplitude gain factors in each polarization, determined separately from the RR and LL data using the same Stokes I source model, yield the desired differential polarization amplitude gain for each antenna throughout the observing run. For the case of spectral-line data, this is only required for the reference antenna at the template scan alone. The R-L gain ratios can also be determined in a fit to direct ratios of the RR and LL data (Kemball, Diamond & Cotton 1995).

In determining R-L amplitude offsets using self-calibration, high signal-to-noise ratio data are required on a compact source, preferably unresolved, as the gain ratio may only be of the order of a few percent for well-calibrated arrays.

Bandpass calibration. The bandpass correction terms $B_m^p(\nu)$ are determined from the parallel-hand data using techniques common to conventional VLBI reduction. The bandpass response functions may be derived from the cross-power or autocorrelation data, although the former is generally recommended. Polarization VLBI reduction is simplified if the bandpass correction does not introduce offsets between the total calibration in each recorded polarization. This is achieved by normalizing the amplitude of the bandpass response function and removing first-order phase slopes across the band by preliminary fringe-fitting if solving for a cross-power bandpass response. The latter step is taken to ensure that the phase response of the bandpass correction contains only terms higher than second-order as a function of frequency and has overall zero-mean phase. It is recommended that the same reference antenna be used in bandpass calibration as will be used in subsequent phase calibration.

Phase calibration. The residual antenna-based instrumental phase ϕ_m^p forms part of the complex antenna gain $G_m^p = g_m^p e^{j\phi_m^p}$. As an antenna-based calibration term, the phase of G_m^p can only be determined relative to a reference antenna, denoted by subscript zero in what follows, and a reference polarization. The residual phase ϕ_m^p is customarily approximated in piece-wise linear form, in terms of residual fringe-rate $\dot{\theta}$ and residual group delay τ, as (Walker 1989),

$$\phi_m^p(\nu,t) = (\nu-\nu_0)(\tau_m^p - \tau_0^p) + (\dot{\theta}_m^p - \dot{\theta}_0^p)(t-t_0) + (\theta_m^p - \theta_0^p) \qquad (25\text{–}2)$$

where (ν_0, t_0) are the reference frequency and reference time of the solution interval. This interval is determined by the phase stability and signal-to-noise ratio of the data.

This and the following section describes phase calibration of polarization VLBI data, as parametrized in Eq. 25–2.

Fringe-fitting. The estimation of residual delays, τ_m^p, and rates, $\dot{\theta}_m^p$, is achieved, as in conventional VLBI, using fringe-fitting techniques. In line with the general calibration strategy outlined here, these quantities can be estimated from the parallel-hand data alone. The utilization of antenna-based global fringe-fitting methods (Schwab & Cotton 1983, Alef & Porcas 1986) removes the need to fringe-fit the weaker cross-polarized data. It is noted, however, that baseline-based methods have been developed to fringe-fit the cross-polarized data, using the parallel-hand fringes to predict their position in $(\tau, \dot{\theta})$ space (Brown et al.1989).

For polarization VLBI data, fringe-fitting can proceed in either multi-band or single-band solution mode. If the pulse-calibration data are available then their use is recommended to align the phases between the recorded basebands, which may have different electronic offsets, and thus increase the available bandwidth and signal-to-noise ratio for multi-band fringe-fitting.

For a circularly unpolarized source model, the delays and rates derived independently from the RR and LL data may be constrained to have the same reference antenna, but will have orthogonal reference polarizations. For the case of Stokes V=0, the offsets between the calibration terms in RCP and LCP can be represented by the R-L delay, rate and phase offsetts at the reference antenna only. These will be denoted by τ_0^{R-L}, $\dot{\theta}_0^{R-L}$ and θ_0^{R-L}, respectively. The offsets can be estimated from the cross-hand data, after application of the parallel-hand fringe-fit solutions. This is equivalent to the use of relations of the form $\tau_m^{R-L} = \tau_m^{RR} - \tau_m^{LL}$ (Brown et al.1989). The RL and LR data for a representative scan on a short baseline yield two independent estimates of the calibration offsets, but with opposite sign, which can be averaged taking this into account. A short baseline is chosen to maximise the signal-to-noise ratio. Leppänen (1995) has demonstrated a method of this nature using a collection of baselines simultaneously.

The contribution of ionospheric Faraday rotation needs to be considered when solving for the calibration offsets (Kemball, Diamond & Cotton 1995), but this factor generally only needs to be taken into account at observing frequencies below approximately 2 GHz. The electronic contribution to $\dot{\theta}_0^{R-L}$ is effectively zero due to the requirements on the phase stability of VLBI electronics. The phase offset θ_0^{R-L} is determined from the cross-hand data and referred to the lowest-frequency baseband. This constitutes a relative phase alignment in which the unknown contribution from the source polarization cancels out. The determination of the absolute R-L phase offset at the reference antenna, or equivalently the absolute electric vector position angle (EVPA), is discussed in Section 2.4.

Self-calibration. Self-calibration allows the determination of the residual phase corrections remaining after fringe-fitting and a simultaneous Stokes I source model, using standard hybrid mapping techniques. In the linearized model for equation (1), these phase corrections can be determined from the parallel-hand data alone, in common with the general approach in this section, where we again assume a circularly unpolarized source model. The offsets between the calibration terms in each sense of recorded polarization, as determined after fringe-fitting, allow the self-calibration phase corrections to be used to calibrate all polarization correlation pairs.

2.3. Feed calibration

At this point in the data reduction sequence, the remaining unknown calibration terms are the instrumental polarization leakage factors D_m^p. In the linearized model, these terms appear only in the cross-polarized correlations, RL and LR. After the calibration steps described above, the remaining unknowns in the expressions for the cross-polarized terms in equation (1) are the leakage factors D_m^p and the linear polarization structure of the source itself. The parallactic angle coefficients are known analytically. The process of separating the source and instrumental polarization is commonly referred to as feed calibration.

All feed calibration strategies for polarization VLBI have in common similar precepts. In the case of adequate signal-to-noise ratio in the cross-polarized data and an observation spanning a sufficient range of parallactic angle, the parallactic angle coefficients are sufficiently non-degenerate to allow the separation of the D^p_m terms and a parametrized representation of the unknown linear polarization structure of the source, such that they can be solved for simultaneously.

The assumptions that may be made regarding the linear polarization structure of the source, or the parametrization that may be used to model the source structure, are the defining characteristic of the different feed calibration techniques. The techniques may be broadly categorized as follows:

- (i) Linearly unpolarized calibrator: $(Q, U) = 0$.
- (ii) Unresolved calibrator: $(Q, U) = \text{const.}$
- (iii) Similarity approximation (Cotton 1993): $Q + jU = \beta I$.
- (iv) Multi-component similarity approximation (Leppänen 1995): $\Sigma_k(Q_k + jU_k) = \beta_k I_k$.
- (v) Spectral line methods (Kemball and Diamond 1997): $Q(\nu) + jU(\nu) = \beta_\nu I(\nu)$.

Finding calibrators that meet condition (i) is sometimes possible in VLBI, particularly at lower observing frequencies, but is rare. Two well-known sources, 3C 84 and OQ208 (Roberts, Brown & Wardle 1991), have frequently been used as unpolarized calibrators at frequencies of 8 GHz or below. Few compact extragalactic radio sources are truly linearly unpolarized, and this approximation may be insufficiently accurate at high frequency or when imaging with high dynamic range.

Case (ii) is relatively infrequent in polarization VLBI, due to the high spatial resolution, although common in connected-element interferometry. The parametrizations adopted in (iii) and (iv) are aimed at dealing with sources that are moderately unresolved or which have relatively simple, compact structure, by modelling the linearly polarizated visibility data as the product of the Stokes I visibility and a complex scaling factor, which is determined in the solution. This is equivalent to fixing the linearly polarized intensity to be a constant fraction of the associated Stokes I intensity, and allowing an arbitrary but constant polarization position angle. In case (iv), the approximation is applied to individual components in the source separately, thus increasing its range of applicability. The component decomposition of the source structure may be specified manually or determined automatically from the aggregation of CLEAN components. No method is well-suited to sources with complex, extended polarized emission, which by definition are poor polarization VLBI calibrators.

The spectral line parametrization (v) allows each frequency channel in the source to form part of an ensemble of polarization calibrators. The independent channels often have unrelated structure, thus reducing systematic error in the estimate of the leakage terms D^p_m. This technique is useful for highly linearly polarized astrophysical masers, which frequently have much higher flux densities than continuum calibrators.

Approximations (ii) through (v) can be iterated to improve both the polarization model for the source and the estimated D^p_m. It is noted that the leakage terms may be frequency-dependent, particularly at lower observing frequencies.

To minimize this effect, leakage terms should be determined in the same frequency range used when observing the target source. This may be particularly important in some spectral line observations. The leakage terms are also time-variable, and may not necessarily be transferrable from one observing session to another.

It is noted that iterative, sequential solutions for G_m^p and D_m^p are possible using the full, coupled system of equations given in (1) in some data reduction packages, including AIPS++. This is useful in general, but specifically for cases where the D-terms exceed 10% in magnitude, and allows an alternative form of polarization self-calibration without explicit parametrization of the source model.

2.4. Absolute EVPA determination

Uncertainty in the EVPA is equivalent to uncertainty in the absolute R-L phase offset at the reference antenna, as noted above. This can be determined with reference to a compact source of known EVPA observed in the same VLBI observing run, and simultaneously measured in polarization using a connected-element array or single dish polarimeter. It is important to note that such sources are often highly variable in linear polarization on short time-scales and secondary observations need to be truly simultaneous to be of value. Note that unless instrumental calibration of the R-L phase offset is possible such secondary observations are frequently referred to a standard source with a component having a known, stable EVPA, such as 3C286 or 3C138.

A small category of calibrators can be used directly in the VLBI observing run to measure the absolute EVPA, thus avoiding the secondary EVPA referencing discussed above. At frequencies of 5 GHz and below 3C 286 or 3C 138 can be used (Cotton et al.1997). At higher frequencies, 3C 279 or 2134+004 (Taylor 1998, G.B. Taylor, private communication) can be used.

If pulse calibration data are applied during phase calibration then the R-L phase offset may be determined from previous measurements, as the R-L pulse calibration offset is expected generally to be stable over extended periods of time, barring electronic modifications at the antenna.

3. Imaging

The calibrated Stokes parameters in the (u, v) plane are derived directly from the corrected polarization correlations, as described above. These can then be imaged using standard synthesis techniques to derive the resultant images in all four Stokes parameters. If the cross-polarized correlations are sufficiently unequally sampled, then full complex imaging and deconvolution of $P = Q + jU$ may be warranted, as described by Conway & Kronberg (1969), Cotton (1993), and Kemball, Diamond & Cotton (1995).

4. Other polarization VLBI observing modes

4.1. Spectral line polarization VLBI

Techniques for spectral line polarization VLBI calibration are discussed by Garcia-Barreto et al.(1988) and Kemball, Diamond & Cotton (1995). In this case it is not always possible to assume that the sources are circularly unpolarized (Stokes V=0). This assumption is incorrect for several astrophysical maser transitions which are often the target of observations in this mode.

In this case, the approach adopted is to initially calibrate the data in a reference polarization (R or L) using standard spectral line VLBI techniques, as outlined elsewhere in this lecture series. An initial correction for parallactic angle is required. The phase offsets between R and L then need to be determined for each antenna individually rather than for the reference antenna alone, in order to transfer the calibration from the reference polarization to the orthogonal recorded polarization. The alignment of the delay solutions, however, which are derived from continuum calibrators requires only the reference antenna polarization delay offset. These sources may be assumed circularly unpolarized, and the discussion follows that in the previous section. The phase offsets at each antenna are derived from (RR,LL) ratios of the continuum calibrator cross-power data, as described by Kemball, Diamond & Cotton (1995).

4.2. Orbiting polarization VLBI

The HALCA antenna, launched by the VSOP mission, is the only active orbiting VLBI antenna at the time of writing. Successful calibration of polarization VLBI observations using the HALCA satellite has been demonstrated by Kemball et al.(1998).

An orbiting VLBI antenna introduces several key differences over polarization VLBI observations with a ground-based array alone. These include the increased resolution of the array, due to the longer baselines, and the lower sensitivity on these baselines, as a result of the smaller effective aperture of the space antenna. This directly impacts the choice and availability of polarization calibrators, which is governed by considerations discussed in the previous section on feed calibration.

Individual space VLBI missions may be subject to more specific constraints. For example, the HALCA antenna is sensitive to only one sense of circular polarization (LCP), and cannot switch pointing direction between target and calibrator sources sufficiently rapidly to allow conventional polarization calibration strategies to be applied. The single polarization does not preclude polarization VLBI observing but does introduce asymetric sampling in RL and LR in the (u, v) plane, thus requiring complex deconvolution (Cotton 1993). The switching cycle time does, however, require that polarization calibration be undertaken using the target source itself, using a method that makes allowance for source structure, as previously discussed. Note that all phase calibration for orbiting polarization VLBI can be referred to a ground-based reference antenna, and can be treated in the same manner as ground-based polarization VLBI calibration in this respect.

A final comment regarding orbiting polarization VLBI concerns the orientation of the orbiting antenna. This may be fixed per pointing, as in the case of HALCA, but in any event will generally not vary in the same manner as ground-based antennas. Even for a fixed orbiting antenna orientation however, there will still be parallactic angle variation on the space baselines to ground-based alt-az antennas, such that conventional polarization VLBI feed calibration strategies can be attempted.

5. Observing and scheduling

In planning polarization VLBI observations, there are several scheduling guidelines which are recommended to simplify subsequent data reduction. The primary guideline is to schedule adequate calibration observations to meet the special requirements of polarization VLBI reduction. This is often the limiting factor in defining the scientific usefulness of a given polarization VLBI observation.

There are several categories for which calibration observations are required, but some of these may overlap in practice. The categories include the following areas of calibration: i) bandpass; ii) amplitude; iii) polarization offsets; iv) instrumental polarization; and v) absolute EVPA. These are discussed individually below.

The choice of bandpass calibrators follows the selection guidelines applicable to conventional VLBI observations. Generally, these sources are selected primarily to yield adequate correlated flux density on the longest baselines in the array, and thus allow a cross-power bandpass solution of acceptable signal-to-noise ratio at all antennas. As they are not imaged, their spatial structure is not of primary concern. The bandpass calibrator may frequently be used also to measure the polarization calibration offsets, such as τ_0^{R-L} and θ_0^{R-L}, as this also primarily requires correlated flux density on short to intermediate baselines and does not depend on source structure. For polarization offset determination, it is beneficial if these sources are linearly polarized at the level of at least a few percent, but this is not always essential as the D-terms themselves will contribute cross-polarized flux density at this level in most cases. Calibrators in this category should be scheduled once every 3-4 hours, with typical scan lengths of ~ 10 minutes. Examples of sources in this category include well-known compact extra-galactic sources such as 3C 273, 3C 279 and 3C 454.3.

Amplitude calibrators are also selected using the same criteria in conventional VLBI. The primary requirement is that these sources be ultra-compact and of known flux density, and most arrays recommend candidate sources for each observing band. Well-chosen amplitude calibrators need only be observed infrequently to set the absolute flux density scale. These sources may typically be observed at the start and end of an observing run. They should be observed at elevations above which the antenna gains are well-known, and also such that atmospheric contributions are minimized.

The choice of polarization calibrator for feed calibration is best considered jointly with the selection of a source for absolute EVPA determination. An unresolved source with sufficient cross-polarized flux density to allow a high SNR fringe-detection (at least 7σ) over a representative fringe-fit solution interval for the observing band, meets both requirements. Sources in this category are often compact BL Lac objects, with examples such as 0235+164 and 1334-127.

Compactness is critical when considering EVPA calibration, but feed calibration, when considered as a separate problem, admits other possibilities. At frequencies below 8 GHz, sources with extended structure but very low linear polarization are often used for feed calibration. Examples in this category include 3C 84 and OQ208. Linearly unpolarized sources do not need to be imaged for feed calibration but adequate observing time should be allocated nonetheless. It is reasonable to allocate a third of an observing run to polarization calibration in general. If a linearly unpolarized source is used for feed calibration, then a separate EVPA calibrator will need to be scheduled. Note also that the similarity methods (Cotton 1993, Leppänen 1995) extend the degree of source structure allowable in the feed calibrator, although weak sources with extended, polarized emission should be avoided in all cases. These methods often allow the target source itself to act as a secondary polarization calibrator. EVPA calibration sources must be observed simultaneously using a connected-element array or single dish polarimeter to determine their absolute EVPA. The on-line catalog maintained by the University of Michigan (http://www.astro.lsa.umich.edu/obs/radiotel/umrao.html) may be consulted for reference in this regard.

Multiple polarization calibrators provide a useful insight into the fidelity of polarization VLBI images, by allowing a direct comparison between the images resulting from the use of different calibrators. The errors in the estimated polarization leakage terms D_m^p are typically dominated by systematic effects due to poorly modelled polarization structure of the calibrator. The impact of D-term errors on the polarization VLBI image depends on the distribution of D_m^p errors within the array, the source structure and parallactic angle coverage. This question is discussed by Roberts, Wardle & Brown (1994) and Leppänen (1995).

In general, feed calibrators should be observed in the same frequency band as the target source to minimize effects due to possible frequency dependence of the D-terms. See Lecture 3 for a discussion of the physical origins of this effect. In this respect, baseband bandwidths no greater than 8 MHz may be recommended at the lower frequencies, if the basebands are to be calibrated separately. It is also critical that the feed calibrators be observed frequently over an adequate range of parallactic angle (at least 100 degrees). This is a fundamental requirement for all polarization calibration.

In summary, polarization VLBI is an accessible technique for all observers, given the good instrumental polarization response of current VLBI arrays and the available calibration software.

References

Alef, W., & Porcas, R.W. 1986, *A&A*, 168, 365,

Bloemhof, E.E., Reid, M.J., & Moran, J.M. 1992, *ApJ*, 397, 500.

Brown, L.F., Roberts, D.H., & Wardle, J.F.C. 1989, *AJ*, 97, 1522

Cawthorne, T. V., Wardle, J. F., C., Roberts, D. H., & Gabuzda, D. C., 1993, *ApJ*, 416, 519.

Conway, R. G., & Kronberg, P. P. 1969, *MNRAS*, 142, 11.

Cotton, W.D., Geldzahler, B.J., & Marcaide, J.M., et al., 1984, *ApJ*, 286, 503

Cotton, W.D. 1989, Polarimetry, in *Very Long Baseline Interferometry: Techniques and Applications*, eds. M. Felli & R.E. Spencer (Kluwer Academic Publishers: Dordrecht), p. 275

Cotton, W.D. 1993, *AJ*, 106, 1241

Cotton, W. D., Fanti, C., Dallacasa, D., Foley, A. R., Schilizzi, R. T., & Spencer, R. E. 1997, *A&A*, 325, 493.

Fomalont, E. B., & Wright, M. C. H. 1974, in *Interferometry and Aperture Synthesis* eds. G. L. Verschuur & K. I. Kellermann, (Springer Verlag: Berlin), p. 256

Garcia-Barreto, J.A., Burke, B.F., Reid, M.J., et al., 1988, *ApJ*, 326, 954

Gabuzda, D.C., Cawthorne, T.V., Roberts, D.H., & Wardle, J.F.C. 1992, *ApJ*, 388, 40

Hamaker, J. P., Bregman, J. D., & Sault, R. J. 1996, *A&AS*, 117, 137.

Kemball, A.J., & Diamond, P.J. 1993, in *Astrophysical Masers*, eds. A.W. Clegg & G.E. Nedoluha, (Springer-Verlag: Berlin), p. 369

Kemball, A. J., Diamond. P. J., & Cotton, W. D. 1995, *A&AS*, 110, 383.

Kemball, A. J., & Diamond, P. J. 1997, *ApJ*, 481, L111.

Kemball, A. J. et al., 1998, *Polarization VLBI observations at 1.6 GHz and 5 GHz with the HALCA satellite: first results from in-orbit checkout observations*, (HALCA memorandum).

Leppänen, K. J., Zensus, J. A., & Diamond, P. J. 1995, *AJ*, 110, 2479.

Leppänen, K. J. 1995, Ph.D. thesis, Helsinki Univ. Technology.

Morris, D., Radhakrishnan, V., & Seielstad, G. A. 1964, *ApJ*, 139, 551.

Reid, M.J., Haschick, A.D., Burke, B.F., Moran, J.M., Johnston, K.J. et al., 1980, *ApJ*, 239, 89

Roberts, D.H., Potash, R.I., Wardle, J.F.C., Rogers, A.E.E., & Burke, B.F. 1984, in: *Proc. IAU Symp. 110, VLBI and Compact Radio Sources* eds. R. Fanti, K. Kellermann, & G. Setti (Reidel: Dordrecht), p. 35

Roberts, D.H., Brown, L.F., & Wardle, J.F.C. 1991, in: *IAU Coll. 131, ASP Conference Series, Vol. 19, Radio Interferometry: Theory, Techniques and Applications.* eds. T. J. Cornwell & R. A. Perley (ASP: San Francisco), p. 281

Roberts, D. H., Wardle, J. F., C., & Brown, L. F. 1994, *ApJ*, 427, 718.

Schwab, F.R., & Cotton, W.D. 1983, *AJ*, 88, 688

Taylor, G. B. 1998, *ApJ*, 506, 637.

Thompson, A. R, Moran, J.M , & Swenson, G.W , 1986, *Interferometry and Synthesis in Radio Astronomy*, (New York: John Wiley and Sons).

Walker R. C. 1989, in: *Very Long Baseline Interferometry: Techniques and Applications* eds. M. Felli, & R. E. Spencer, (Kluwer Academic Publishers: Dordrecht), p. 141

Wardle, J. F., C., & Roberts, D. H. 1994, *in Compact Extragalactic Radio Sources* eds. J. A. Zensus & K. I. Kellermann (Socorro: NRAO), 217.

Weiler, K. 1973, *A&A*, 26, 403.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

26. Space Very Long Baseline Interferometry

J. S. Ulvestad
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. Space Very Long Baseline Interferometry (SVLBI) is a technique in which an array of ground radio telescopes observes a source in conjunction with one or more orbiting radio telescopes. A synthesized aperture is produced with a diameter approximately equal to the apogee of the spacecraft orbit, enabling higher resolution imaging than is possible with ground-only VLBI. This paper discusses the characteristics and techniques of SVLBI, emphasizing the differences in capabilities and data-processing from traditional VLBI. General concepts are illustrated with a number of examples from the currently operational VLBI Space Observatory Programme, using a dedicated SVLBI satellite that was launched in February 1997.

1. Introduction

Space Very Long Baseline Interferometry (hereafter SVLBI) is a technique that is used to extend the maximum baseline available in VLBI observations. This is done by launching a radio telescope into space, preferably in an elliptical orbit around the Earth. The SVLBI satellite then observes in conjunction with ground radio telescopes, synthesizing an aperture whose effective resolution is that of a radio telescope much larger than the Earth. This enables imaging of the most compact radio sources in the universe with sub-milliarcsecond resolution.

The first SVLBI demonstrations were carried out from 1986 through 1988 (Levy et al. 1986, 1989; Linfield et al. 1990), using an element of NASA's Tracking and Data Relay Satellite System (TDRSS). A 4.9-meter radio antenna aboard a geostationary satellite above the western Atlantic Ocean was used as the orbiting telescope. Successful observations were carried out at observing frequencies of 2.3 and 15 GHz, demonstrating that there were no insurmountable technical obstacles to scientific observations with a dedicated SVLBI satellite. In the most extensive series of observations, carried out in early 1987, a number of sources were shown directly to have apparent rest-frame brightness temperatures in excess of 10^{12} K (Linfield et al. 1989).

The success of the original demonstrations was an important argument for the proposed QUASAT (Schilizzi et al. 1984) and IVS (International VLBI Satellite; see Pilbratt 1991) concepts, neither of which was ultimately approved. However, the TDRSS demonstrations led to development of the HALCA (Highly Advanced Laboratory for Communications and Astronomy) satellite of the VLBI Space Observatory Programme (VSOP). HALCA was developed by the Institute of Space and Astronautical Science (ISAS) in Japan, and launched in February 1997. It carries an 8-meter radio telescope that works with ground telescopes to image radio sources at 1.6 and 5 GHz (see Hirosawa & Hirabayashi 1995; Hirabayashi 1998; Hirabayashi et al. 1998). VSOP is used for many examples in this article; more details can be found in the VSOP Proposer's Guide (VSOP Science Operations Group 1998). Another mission, RadioAstron (Kardashev 1997), awaits a launch some time after the turn of the century.

2. Orbit Influence on (u,v) Coverage

The primary purpose of doing SVLBI is to increase the maximum baseline length sampled in a given observation. However, baselines that are too long relative to the ground baselines can give large holes in the (u,v) plane, leading to significant imaging defects. Furthermore, evolution of the orbit parameters leads to changes in the (u,v) coverage over time scales of months, having a significant impact on observation planning. This section discusses some of the effects of the orbit on SVLBI observations.

2.1. Orbit Parameterization

Typically, a spacecraft orbit is defined by six fundamental parameters (see Bate, Mueller, & White 1971, p. 59, for a useful sketch):

- a = semi-major axis of the orbit.
- e = orbital eccentricity.
- i = inclination with respect to the Earth's equator.
- Ω = longitude of the ascending node. This is the angle in the Earth's equatorial plane, in the same sense as the motion of the satellite, from the direction of the vernal equinox to the ascending node. The ascending node is the intersection of the equatorial plane with the satellite's orbital plane, at the location where the satellite moves north of the equator.
- ω = argument of perigee. This is the angular distance from the ascending node, along the orbit plane, to the direction of perigee.
- T = time of the satellite's perigee passage.

The location of the satellite at a particular time is often represented by the mean anomaly, M. This gives the fraction of an orbital period (often expressed in degrees) traversed since the last perigee passage. For example, for a spacecraft in an orbit with an 8-hr period, which passes perigee at 02:00 UT, $M = 135^\circ$ at 05:00 UT and $M = 180^\circ$ at 14:00 UT.

2.2. Orbit Selection

Selection of an orbit is akin to choosing the locations of the telescopes in design of a synthesis array such as the VLA or the VLBA (e.g., Walker 1984). One might think that the top priority is to launch the spacecraft into the highest possible orbit. However, very high orbits have several disadvantages. First, large holes are created in the (u,v) plane, limiting the dynamic range of images. Second, the orbit size should be scaled to the size of the physical phenomena of interest. If the orbit is too high, source brightnesses may be too low for detection, or sources may vary during a single orbit (a suitable imaging interval). Third, launch vehicles are usually a substantial fraction (10%–30%) of the overall mission cost of hundreds of millions of dollars, so budget limitations constrain the apogee height. Finally, for fixed properties of the data transmission system, the maximum downlink data rate (hence observing bandwidth) is reduced according

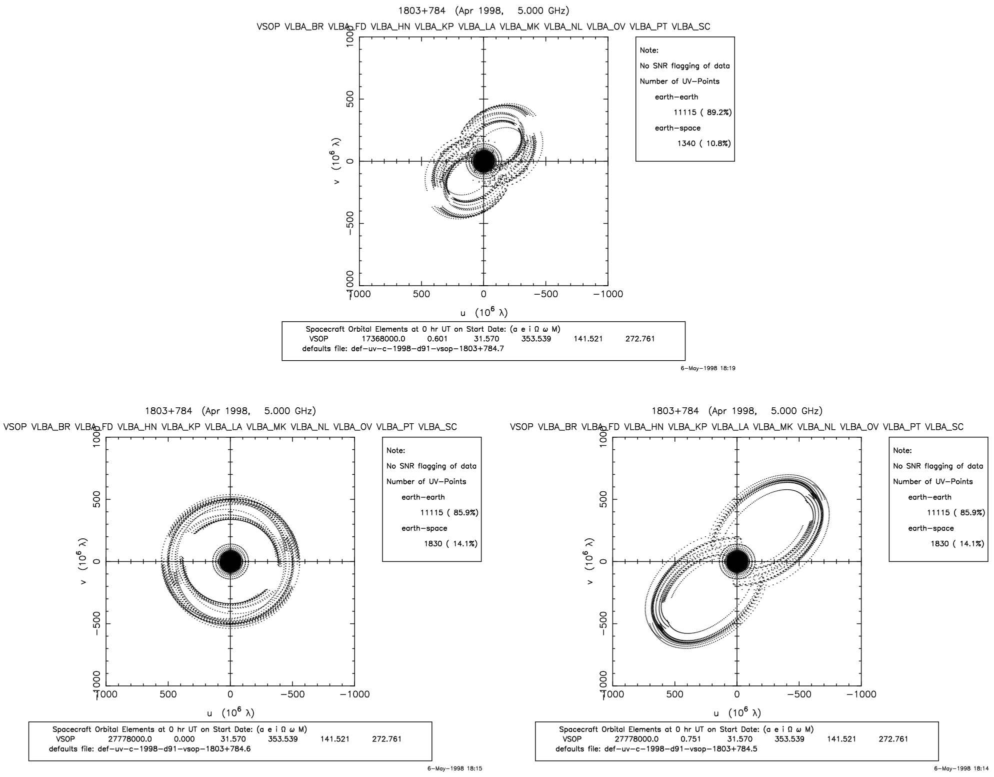


Figure 26–1. Sample 5-GHz SVLBI (u,v) coverages for 24-hr (spacecraft + VLBA) observations of the quasar 1803+784 on 1 April 1998. All plots are at the same scale, from -10^9 to 10^9 wavelengths in u and v. *Top:* HALCA orbit. *Lower left:* Circular orbit with height equal to HALCA apogee height. *Lower right:* Highly elliptical orbit with semi-major axis equal to the radius of the circular orbit in the lower left panel.

to the inverse-square law. If the orbit is too high, the correlated flux will drop and the observing bandwidth must be decreased, seriously degrading the signal-to-noise ratio for fringe detection.

Figure 26–1 illustrates some simple consequences of orbit selection for a hypothetical 24-hr observation of the northern quasar 1803+784 on 1 April 1998, as simulated using FAKESAT (Murphy et al. 1994; Murphy 1995). The top panel shows the (u,v) coverage for an observation using the VLBA with HALCA, whose orbital elements on that date were $(a, e, i, \Omega, \omega, T)$ = (17,368 km, 0.601, 31.57°, 353.4°, 141.5°, 01:32 UT on 1998 April 1). This elliptical orbit has an apogee height of 21,400 km, a perigee height of 540 km, and a 6.3-hr period. The lower left panel shows the (u,v) coverage for a circular orbit with a height of 21,400 km. This orbit has $a = 27,778$ km and $e = 0$ (12.8-hr period). The maximum baseline length is similar to that for HALCA, but there is a large gap between the ground baselines and the longer space-ground baselines. Finally, the lower right panel shows the coverage for a larger elliptical orbit, also having $a = 27,778$ km and a 12.8-hr period. Here, the perigee height is the same as HALCA, 540 km, but the apogee height is 42,300 km, and $e = 0.751$. For this orbit, the longest baseline has been increased considerably, and the shortest ground-space baselines overlap the ground-ground baselines, but the size of the holes in the (u,v) plane is quite large.

2.3. Variations in u and v During Observation

The imaging of interferometry data is done by gridding the sampled visibilities onto the (u,v) plane and then Fourier transforming them to create an image (see Lecture 7). The classic problem of time-average smearing is caused by assigning values of u and v corresponding to the midpoint of an averaging interval, when they actually vary throughout that interval (see Lecture 18). Since correlator integration times are usually about a second, and SVLBI fields of view are generally small, time-average smearing is avoidable unless the data set must be averaged to make computations practical. (For details, see VSOP Science Operations Group 1998, p. 25.)

A related problem is caused by the $\sim$ 10 km s^{-1} spacecraft speed near perigee, combined with the fact that the projected space-ground baselines may be as short as $\sim$ 1000 km at the same time. Integration times of hundreds of seconds often are required to detect fringes to the spacecraft. This implies that u and v can double during a fringe-fit interval, so the structure phase in the visibility function may change substantially, causing a loss of coherence (and signal-to-noise ratio) in the fringe fit on time scales longer than tens of seconds. The effect is much smaller near apogee, since u and v are considerably larger, and the spacecraft is moving only $\sim$1–3 km s^{-1}.

2.4. Orbit Effects on (u,v) Coverage in Different Directions

The resolution of an interferometer depends on its baseline length projected on the plane normal to the source direction. In SVLBI, the longest baselines lie in the spacecraft orbit plane. Therefore, the best two-dimensional (u,v) coverage will occur for sources lying in directions near the normal to the orbit plane, while the coverage for sources near that plane will be extremely elongated. To illustrate this point, Figure 26–2 is a FAKESAT (Murphy 1995) plot of the (u,v) coverage as a function of position on the sky for 24-hr observations made with HALCA and the VLBA, on 1 April 1998. For simplicity, and to illustrate the basic points, all observing constraints, including the Sun-avoidance angle (see Section 3.6.), have been neglected. In the figure, the sinusoidal line with an amplitude of 31.6° is the plane of the spacecraft orbit, P and A indicate the perigee and apogee directions, respectively, and N represents the two normals to the orbit plane. (A Sun angle of 70° also is shown by the curved lines centered on the Sun location.) Inspection of the figure shows that the (u,v) coverage is basically one-dimensional near the orbit plane and two-dimensional near the orbit normals, and that the lengths of the projected baselines are shortest for sources in the directions near apogee and perigee. Sources in the far south are not visible to the VLBA, so there is no (u,v) coverage at $-80°$ declination.

2.5. Effects of Orbit Precession on (u,v) Coverage at Different Times

In Figure 26–2, equatorial sources near 12 hours right ascension have (u,v) coverage that is quite one-dimensional. Two very important VLBI sources, 3C 273 and 3C 279, are in this vicinity, so it seems that VSOP would be a poor imaging mission for these quasars. However, the geometry changes because

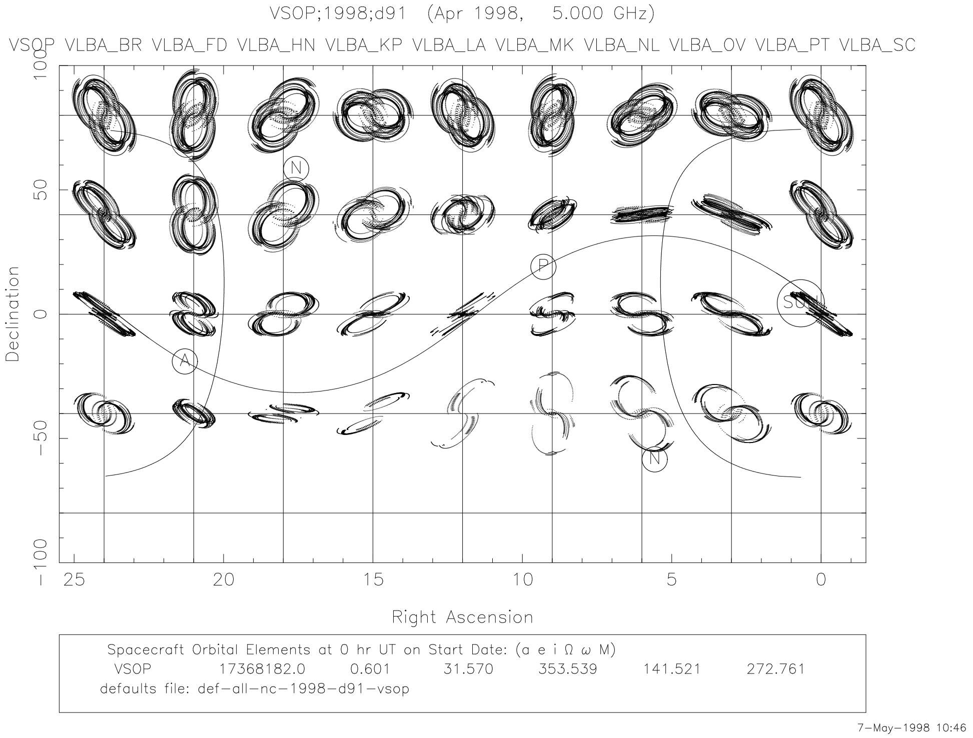


Figure 26–2. 5-GHz SVLBI (u,v) coverages as a function of sky position, for HALCA+VLBA, on 1 April 1998. No HALCA observing constraints are included other than the requirement that there be a line of sight between the spacecraft and a tracking station.

of the Earth's oblateness, particularly the J_2 term in the Earth's gravitational potential, which causes both Ω and ω to precess. (Precession caused by the Moon and Sun are generally negligible, by comparison.) This changes both the orientation of the orbit plane and the location of perigee within that plane. Thus, precession improves the (u,v) coverage for a given source at some epochs, but the non-repeating coverages make the interpretation of monitoring observations much more difficult.

The equations for the rates of precession can be found in many references (e.g., Griffin & French 1991; Boden 1992); the precession rates due to J_2 are

$$\dot{\Omega}(J_2) = -2.065\ (a/10,000\ \mathrm{km})^{-7/2}\ (\cos i)(1-e^2)^{-2}\ \mathrm{deg/day} \qquad (26\text{–}1)$$

and

$$\dot{\omega}(J_2) = 1.032\ (a/10,000\ \mathrm{km})^{-7/2}\ (4-5\sin^2 i)(1-e^2)^{-2}\ \mathrm{deg/day}\,. \qquad (26\text{–}2)$$

For the HALCA orbit, $\dot{\Omega}(J_2) \approx -0.62$ deg/day and $\dot{\omega}(J_2) \approx 0.96$ deg/day. Thus, the precession periods for Ω and ω range between 1 and 1.7 years, less than the mission lifetime of several years. This implies good imaging capability for most

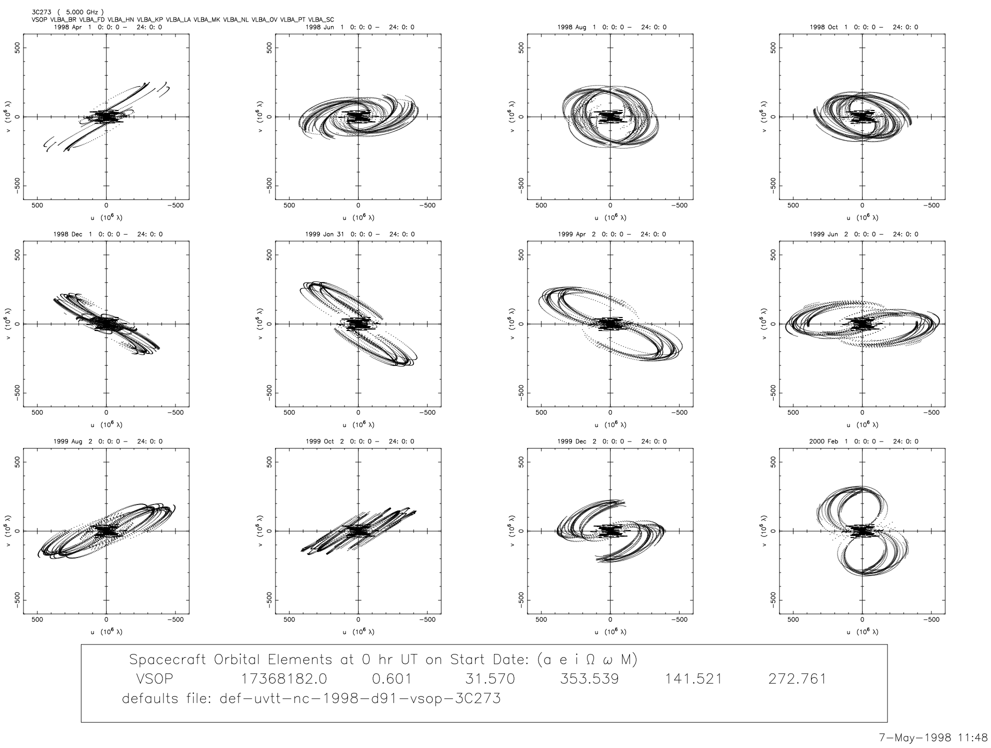


Figure 26–3. 5-GHz SVLBI (u,v) coverages of 3C 273 at two-month intervals, for HALCA+VLBA, starting on 1 April 1998. No HALCA observing constraints are included other than the requirement that there be a line of sight between the spacecraft and a tracking station.

sources at some time during the VSOP mission. It is useful to note that $\dot{\omega} = 0$ if $i = 63.4°$, which can be used to maximize tracking time by keeping perigee in the south (see Section 3.3.).

Figure 26–3 is a FAKESAT (Murphy 1995) plot of the (u,v) coverage for 3C 273 as a function of time. The coverage is plotted at two-month intervals beginning in April 1998, for an array consisting of HALCA and the VLBA. As in Figure 26–2, all observing constraints have been suppressed in order to highlight the effects of orbit precession. The plot shows epochs when the coverage is nearly one-dimensional (source very close to the orbit plane), epochs when the coverage is rather two-dimensional with fairly short baselines (source close to perigee or apogee directions), and epochs where the coverage is somewhat two-dimensional with longer baselines. There is never completely two-dimensional coverage with the longest baselines, since 3C 273 is never more than 35° from the orbit plane.

3. The Space Radio Telescope

3.1. What is the "Space Radio Telescope"?

One simple way of thinking of a VLBI telescope is that it is a system consisting of five major elements or subsystems:

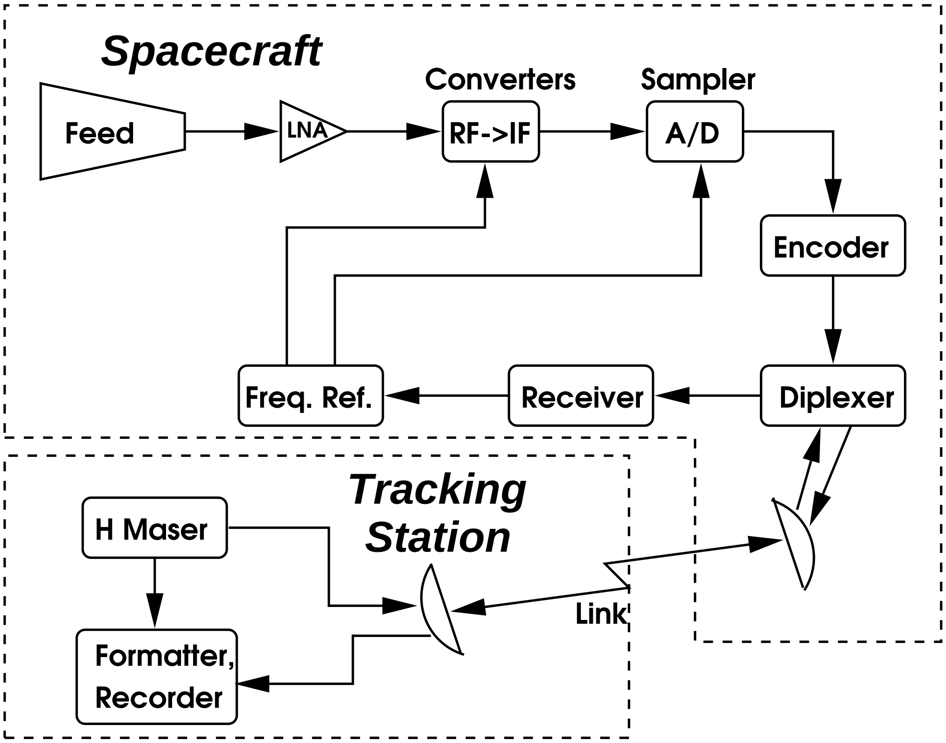


Figure 26–4. Simple block diagram of a generic Space Radio Telescope, consisting of an orbiting VLBI satellite and a tracking station.

- Observing antenna, feed, and receivers
- Frequency converters, filters, and IF data processors
- Digitizers and samplers
- Time and frequency standard
- Formatter and wideband VLBI data recorder(s)

At ground observatories, the elements listed above are co-located, within a few tens of kilometers or less. However, in SVLBI, some subsystems are located on the spacecraft, while others are at a ground tracking station. For example, in the TDRSS experiments, only the first two subsystems were in space, while the others were located at the White Sands Ground Terminal. For VSOP, the first three subsystems are aboard HALCA, while the frequency standard and data recorders are at the tracking stations. Future missions may employ space-borne hydrogen maser clocks, although ground-based frequency standards still may be required for a variety of purposes. The combination of the five fundamental subsystems at the orbiter and the tracking station(s) is defined here as the "Space Radio Telescope." A simple block diagram, including additional elements required to connect the spacecraft to the ground, is shown in Figure 26–4.

Wideband VLBI data cannot be accumulated and stored aboard a spacecraft during an observation, so they must be recorded on the ground in real time. Therefore, good sampling of the (u,v) plane requires a network of tracking stations distributed around the Earth. At a given time, the Space Radio Telescope may be defined to be the orbiter and a particular tracking station, while the Space Radio Telescope for an entire scientific observation might be defined as the combination of the spacecraft and all the tracking stations that participate in that observation. The difference between the two definitions is not fundamental; the important point is that there is no operational Space Radio Telescope unless the orbiter is observing a source *and* has a reliable link with a tracking station.

3.2. The Spacecraft

The SVLBI orbiter is the "front end" of the Space Radio Telescope. General properties of the spacecraft (mass, power, attitude control, etc.) can be found elsewhere (e.g., Kardashev 1997; Hirabayashi 1998; VSOP Science Operations Group 1998; Hirabayashi et al. 1998). The critical element of the science payload is the receiving antenna, which typically works at centimeter wavelengths and is up to 10 meters in diameter; details of the design of the HALCA antenna can be found in Natori et al. (1994) and Takano et al. (1994). The receivers and IF processing subsystem are also necessary ingredients aboard the spacecraft, but are not described here, since they are fundamentally similar to the subsystems found at ground observatories. Cooling of the receivers has not been practical to date due to mass and power considerations, but is planned for the second generation of missions, VSOP-2 and ARISE (see Section 6.).

3.3. Tracking Stations

Tracking stations are necessary to record the VLBI data and (so far) to provide the stable frequency reference for the spacecraft. In the case of VSOP, five tracking stations are employed, in California, West Virginia, Spain, Japan, and Australia. Four are located in the northern hemisphere, so the tracking and (u,v) coverage are much better when HALCA has apogee in the north ($0° < \omega < 180°$), and somewhat degraded when the orbit has precessed so that apogee is in the south.

The primary functions of the tracking stations are to provide a stable frequency reference, derive the correct time at which data samples were received aboard the spacecraft, provide the data necessary for accurate orbit determination, and record the wideband VLBI data. Details of the design of a tracking station are beyond the scope of the present article, but some relevant information can be found in D'Addario (1991) and in Kawaguchi et al. (1994). A separate "command" station also may be used to send commands and receive housekeeping telemetry from the spacecraft.

3.4. Timing and Data Link

In ground VLBI observations, the clock, recorder, and receiving antenna are co-located, making it relatively straightforward to time-tag the data. However, in the Space Radio Telescope (see Figure 26–4), this is not the case. The TDRSS experiments pioneered a method of "time transfer" (or "phase transfer") in which a ground frequency standard was used as the reference for the VLBI subsystems in space (e.g., Levy et al. 1989).

Many details of time transfer in orbiting VLBI are described by D'Addario (1991), and are only summarized here. On the ground, a pure frequency tone referenced to a hydrogen maser clock is generated and uplinked to the spacecraft. A predicted spacecraft orbit is used to calculate the Doppler shift, and the emitted frequency of the uplink tone is varied to cause a constant frequency to be received at the spacecraft. There, the tone is received, detected, and locks a frequency reference that is used to operate the on-board oscillators. The tone also is transponded to a different frequency and re-transmitted to the ground tracking station. On the ground, the received tone is mixed with the frequency expected for the predicted Doppler shift. The residual frequency (received minus expected) is measured by a quadrature phase detector at a high rate, and the phase residual is recorded for later use.

If the predicted Doppler shift for the spacecraft is error-free, and there are no other signal delays, the phase residual will be zero, implying that the spacecraft clock was running at the correct rate. However, errors in the predicted orbit cause the residual to be nonzero. This implies that the spacecraft clock rate was in error, and that the digital data were sampled at incorrect intervals, which must be corrected in data correlation (see Section 4.2.). Only the uplink (one-way) error affects the spacecraft timing, but the tracking station measures a two-way residual, which must be converted to the time error for the spacecraft.

The digitized VLBI data are downlinked in a series of telemetry frames, each led by a header containing auxiliary information about the spacecraft and its science payload. These wideband data are demodulated on the ground and recorded on digital tape in a format similar to that used by the ground radio telescopes. (In order to prevent spurious correlations, the header data are replaced by pseudo-random noise before recording.) The rate of data recording is derived from a clock driven by the downlink telemetry, not by the ground hydrogen maser. This "data clock" runs at a variable rate in the ground frame, depending on the Doppler shift imposed by the spacecraft motion. The absolute offset of the data clock from the ground clock also must be measured at some initialization epoch, to enable VLBI correlation.

The example of VSOP is the best way to illustrate the activities necessary for correct data recording and timing. A (more-or-less) chronological sequence of activities is listed below:

- Tracking station uplinks a tone to the spacecraft, with frequency varied so as to achieve a received frequency of exactly 15.3 GHz aboard HALCA.

- Spacecraft receives uplink tone and uses it to lock all local oscillators.

- VLBI data are sampled and digitized on the spacecraft at 128 Mbit sec^{-1} (two 16-MHz channels, each with 2-bit sampling), as controlled by the clock running at a constant rate in HALCA's rest frame.

- Received frequency is transponded to 14.2 GHz in the spacecraft frame, and re-transmitted to the ground tracking station along with the wideband VLBI data, which are divided into frames 80 kilobytes in length.

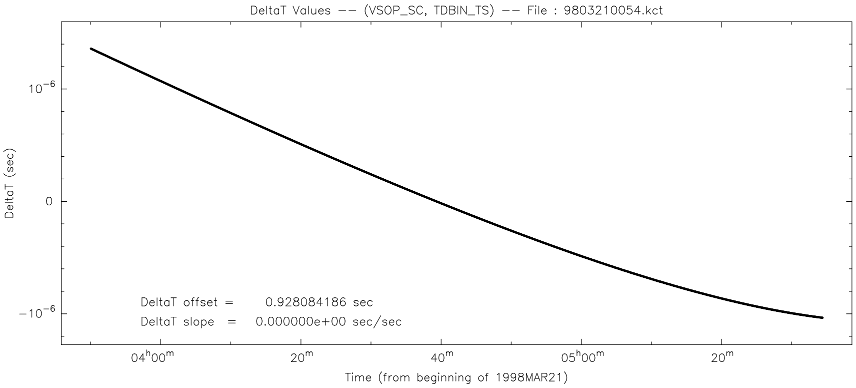


Figure 26–5. Time Corrections File for Tidbinbilla tracking of HALCA, from 03:50 to 05:35 UT on 21 March 1998. The range of corrections is slightly larger than ±1 μsec.

- Tracking station detects the downlink signal, extracts the Doppler-shifted tone, mixes it with the expected received frequency, and measures the residual phase hundreds or thousands of times per second.

- Tracking station detects valid telemetry and sets the VLBI formatter clock. Header data are diverted into a separate data file for later use. For 4 of the 5 VSOP tracking stations, the clock is set to an integer second at the start of a telemetry frame. Offset of the formatter clock from the ground hydrogen maser clock is recorded for future reference.

- Wideband data are demodulated from the downlink and recorded on a VLBI tape for supply to the designated correlator. After initialization, the clock rate is governed by the data clock, with 128 Megabits defining one "second" (differing from a ground "second" by the Doppler shift).

- Two-way phase residuals are recorded throughout the tracking pass, at rates of 400 Hz or higher.

After a tracking pass, a "Time Corrections File" is generated, containing the information needed to relate the time on the VLBI tape to the time that a data sample was received aboard HALCA. This file incorporates the following:

- Clock initialization offset.

- Timing error derived from the measured round-trip phase residual.

- Additional delays in the tracking station hardware.

- Geometric delay (light travel time) on the downlink from spacecraft to tracking station.

- Any other effects considered significant at the tracking station.

The Time Corrections File is supplied to the VLBI correlators for use in the final correlation (see Section 4.2.). Figure 26–5 shows a portion of the Time Corrections File from the Tidbinbilla (Australia) tracking station for a 5-GHz VSOP observation of 3C 345, on 21 March 1998. In this plot, a constant delay offset has been removed. The variation in the clock error over 1.75 hr is about 2.4 μsec, more than 10,000 cycles at the observing wavelength of 6 cm. (The speed of light, in convenient units, is 30 cm nsec^{-1}, so a 1-nsec shift corresponds to 5 wavelengths.) If this clock correction was not applied in correlation, the coherence would be severely degraded (see Section 4.4. and Figure 26–8).

3.5. Orbit Determination

Orbit determination for SVLBI uses the same data from which the time corrections are derived (see Section 3.4.). These data are measurements of the two-way Doppler shift, which can be used to derive the spacecraft orbit. (In addition, range and Doppler measurements may be made from the command station.) The spacecraft orbit is derived using a detailed model of the Earth's gravitational field as well as a solar-pressure model for the spacecraft, taking into account the pointing direction of the large radio antenna. Uncertainties in the gravitational field are especially important if the spacecraft has a low perigee height. Solar-pressure effects are more important than for most spacecraft, because the large radio telescope leads to a high area/mass ratio for the spacecraft.

For HALCA, the orbit determination typically provides reconstructed orbits for VLBI correlation that have r.m.s. position and velocity accuracies of about 15 m and 6 mm sec^{-1}, respectively. These accuracies are good enough for VLBI correlation, but not for global astrometry. Orbits predicted a few days in advance are typically several times less accurate. Second-generation SVLBI missions will make use of Global Positioning System (GPS) receivers aboard the spacecraft, providing position accuracies predicted to be better than 10 cm over the entire orbit.

3.6. Observing Constraints

There are a variety of constraints on observing with a Space Radio Telescope that differ from a ground telescope. There are Sun-angle constraints caused by the necessity to point the spacecraft solar panels at the Sun, requirements to shade parts of the spacecraft from direct sunlight, and limited reaction-wheel capacity to remove torques imposed by solar radiation pressure. Communication constraints include the requirement that a tracking station be able to view the spacecraft, as well as the requirement that the spacecraft telemetry antenna have a clear line of sight to that tracking station. The latter requirement may be violated because of blockages on the spacecraft that depend on observation direction. In addition, there are eclipse constraints, since power and thermal requirements normally will prevent observations during (or near) times when the Sun is eclipsed by either the Earth or the Moon. Finally, there are miscellaneous constraints related to issues such as command-storage capability and spacecraft slew rates. Slew-rate limitations, for instance, make phase-referencing VLBI observations (Beasley & Conway 1995) difficult or impossible for any realistic spacecraft design.

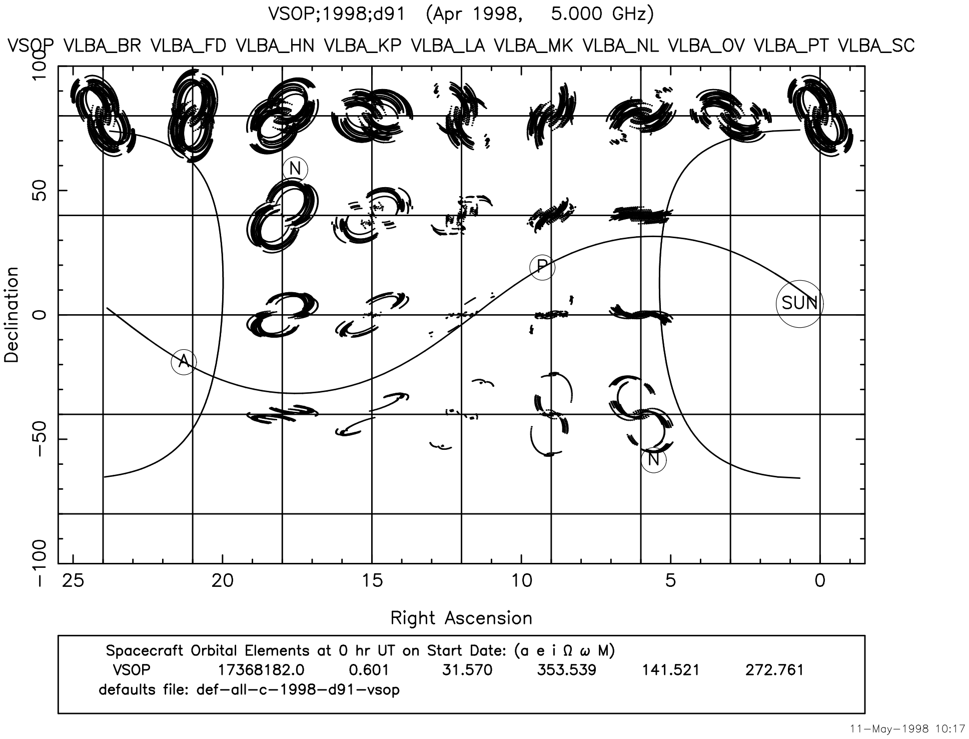


Figure 26–6. 5-GHz SVLBI (u,v) coverages as a function of sky position, for HALCA+VLBA, on 1 April 1998, including all observing constraints on the Space Radio Telescope. (Compare to Figure 26–2.)

Descriptions of VSOP constraints in some detail can be found in the VSOP Proposers Guide (VSOP Science Operations Group 1998); their effect is illustrated in Figures 26–6 and 26–7. Figure 26–6 is the all-sky (u,v) coverage for HALCA and the VLBA on 1 April 1998, similar to Figure 26–2, but now including all relevant constraints on VSOP observing. Note the large exclusion zones in Figure 26–6 due to source proximity to the Sun. Similarly, Figure 26–7 is a repetition of Figure 26–3, a plot of the (u,v) coverage for 3C 273 every two months, with all constraints now included. There are long periods for which no SVLBI observations of 3C 273 are possible, because the source is near the ecliptic plane and is too close to the Sun for about five months per year. These constraints and the variable (u,v) coverage make SVLBI scheduling fairly complex (e.g., Meier 1998).

3.7. Data Calibration

Calibration of the Space Radio Telescope is somewhat different from a ground telescope, and depends on downlink telemetry from the orbiter (see Section 3.4.). Because of command-storage constraints, it may be impossible to fire a noise diode regularly. Also, ground telescopes often incorporate several different noise diodes of differing strengths, but power and mass constraints may prevent this for a spacecraft. For example, aboard HALCA, the only available noise diode

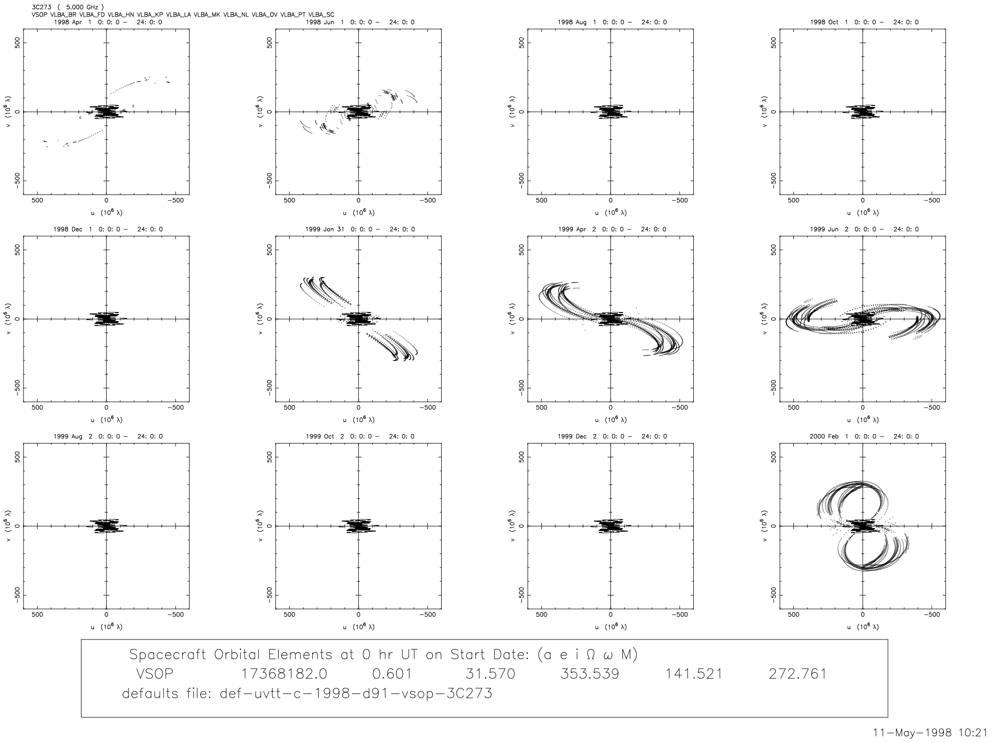


Figure 26–7. 5-GHz SVLBI (u,v) coverages of 3C 273 at two-month intervals, for HALCA+VLBA, starting on 1 April 1998, including all observing constraints on the Space Radio Telescope. (Compare to Figure 26–3.)

approximately doubles the system noise temperature. Therefore it is not beneficial to use the noise source during VLBI observations, since it would significantly increase the system noise temperature.

The low sensitivity and constraints on observing directions also make it difficult to calibrate the pointing using observations of strong radio sources. This can be especially troublesome because the pointing offsets may change substantially as a function of Sun angle. For HALCA, the uncertain pointing calibration made it difficult to attempt fringe detection with the 22-GHz system, which was damaged during launch and has extremely poor sensitivity. (These fringe searches were ultimately successful thanks to a huge flare in the Orion H_2O maser.)

A great advantage in calibration of the Space Radio Telescope is the lack of an atmosphere above the spacecraft. For HALCA, the system temperature shows good repeatability from orbit to orbit, at the level of about 2%. Variations within an orbit may be as large as 5% due to changing thermal conditions near eclipses. However, even these changes are relatively consistent for adjacent orbits, so they can be calibrated by using measurements made on orbits before or after the actual observations. In fact, experience has shown that the system temperature calibration for HALCA is considerably more accurate than for some of the ground radio telescopes involved in VSOP observations.

Polarization calibration of SVLBI data is also rather difficult due to the low signal-to-noise ratio. HALCA observes only in a single polarization (left circular), and the lack of rotation of the parallactic angle makes calibration of the polarization leakage terms rather involved. However, such calibration has been carried out successfully by the VSOP Polarization Study Team as part of in-orbit checkout (Kemball et al. 1998). Polarization calibration for missions with dual polarization and higher sensitivity might be easier, but the off-axis antenna designs under consideration for some future missions will lead to additional complications.

4. Correlation and Data Processing

The VLBI correlation process is described in some detail by Romney (1995) and in Lecture 4, with the actual implementation at the VLBA correlator discussed by Benson (1995). Those discussions are not repeated here. Instead, we concentrate on the fundamental differences and additions that are necessary for SVLBI.

4.1. Delay Model

VLBI correlation requires an accurate delay model for each instant of an observation. The delay model is a representation of the apparent delay in the wavefront received at a radio telescope, typically referred to its arrival time at the Earth's center, and includes many different effects on the signal, as well as the models of the observatory clocks. The data are referenced to the delay model so that the interference fringes are "stopped," and the residual delays and delay rates are small. Small residuals enable long coherent integrations in fringe-fitting, which translates into a capability for detecting fringes on weaker sources.

In ground VLBI, the geometric part of the delay (τ_g) is always negative (the telescope is always "in front of" the Earth's center, as viewed from the radio source), and its maximum magnitude is about 21 msec. The maximum rate of change of the geometric delay is ≤ 3 μsec sec^{-1}. This "fringe rate" or "fringe frequency" often is expressed in units of cycles of the observing frequency, by multiplying the delay rate by the observing frequency. For an observing frequency of 100 GHz, therefore, the maximum ground-VLBI fringe rate is ≤ 300 kHz.

For SVLBI, the maximum delay and fringe rate that must be accommodated are much larger. For example, if the spacecraft has an apogee height of 90,000 km, the delay of the downlink will be approximately 0.3 sec, and the geometric delay may be of the same magnitude, with either the same or the opposite sign. Furthermore, initialization offsets of the station clock relative to ground UTC (see Section 3.4.) also can be on the order of a second. Both the spacecraft geometric delay and the total delay, therefore, can be either positive or negative, and have magnitudes of seconds. Fortunately, most modern correlators can accommodate unlimited delay by offsetting the tapes during playback.

The delay rate (and delay "acceleration") are more difficult to address. A rough approximation to the maximum delay rate can be made by taking the maximum spacecraft speed in an elliptical Earth orbit, ~ 10 km sec^{-1}, and dividing this by the speed of light to yield a delay rate of ~ 33 μsec sec^{-1}. When Earth rotation and slightly higher speeds for an extremely elliptical orbit are included, the correlator needs to accommodate a delay rate approaching 40 μsec sec^{-1}. The delay models inside VLBI correlators are usually expressed as polynomials in time that are good for a limited time interval, and are further approximated as linear functions over a shorter interval (see Benson 1995). Therefore, the higher SVLBI delay rate may require more accurate polynomial fitting and evaluation, more frequent model updates, and even additional hardware.

4.2. Correlator Inputs

Achievement of an accurate delay model requires that the observation geometry and the clock models be well specified, so the correlator can correctly process the digitized VLBI data. For a ground telescope, inputs include an accurate telescope location, a precise Earth-rotation model (including polar motion), an accurate radio-source position, and a measurement of the observatory clock error relative to some fiducial UTC reference, such as the network of GPS satellites. These models and measurements are relatively simple to input, consisting of just a few numbers.

For SVLBI, there are two crucial complications. First, the position of the telescope is not tied to the Earth's surface, but is a much more complicated function. A detailed orbit ephemeris is typically provided as a six-dimensional (position and velocity) spacecraft state vector as a function of time; this ephemeris is used to generate the geometric delay model. Second, the clock model must be represented by a Time Corrections File (see Section 3.4.). This file contains the information necessary to correct for clock errors imposed by the process of "constructing" the Space Radio Telescope. For VSOP, the Time Corrections File is supplied as a series of 10 corrections per second, which are added to the time on the VLBI tape in order to specify the time at which a data sample was acquired. The high frequency of these corrections was specified to guard against changes in the atmosphere or electronics at frequencies as high as 5 Hz, which would cause coherence loss if not properly sampled. Actual experience has shown that fitting a high-order (5th order in time) polynomial to the Time Corrections File every 10 seconds causes no significant loss of coherence. Use of such a polynomial makes correlator implementation simpler, since it can be added to the other components of the delay polynomial (with the appropriate sign!) in generating the correlator model.

Two important input parameters for VLBI correlation are the window sizes for the output residual delay and fringe rate. Standard parameters for ground VLBA observations are about ±4 μsec for the delay, and ±0.25 Hz for the fringe rate. The r.m.s. position error of HALCA (see Section 3.5.) corresponds to less than 100 nsec in delay, so once the constant offsets are determined for a tracking station, it is possible to correlate spacecraft data without increasing the residual delay window. A velocity error of ±1 cm sec^{-1} corresponds to a fringe-rate error of ±0.17 Hz for an observing frequency of 5 GHz, not much smaller than the typical ground-only fringe-rate window. Therefore, at the VLBA correlator, the space-ground baselines of a 5-GHz VSOP experiment are typically output with a 1.04-sec integration time, corresponding to a residual rate window of ±0.48 Hz. This increases the output data rate significantly; in order to keep the sizes of output data sets manageable, SVLBI correlators typically have (and use!) the capability of outputting the ground-ground baselines less frequently than the space-ground baselines.

4.3. Fringe-fitting

Fringe-fitting of SVLBI data has some important distinctions from the process used for ground VLBI data. First, the sensitivity of the Space Radio Telescope is quite low compared to most ground telescopes. Second, because the spacecraft cannot slew rapidly, it is not possible to observe strong calibration sources, so the program source provides the only option for locating the space-ground fringes. Third, location of fringes for the Space Radio Telescope is dependent on the accuracy of the orbit and of the Time Corrections File, so the initial searches often must employ wider search windows than for ground VLBI.

In order to assess the likelihood of detecting fringes, and the limitations on the source strength that can be detected for a given SVLBI mission, we use the following formula for the r.m.s. noise (ΔS_{ij}) for a weak source on a single baseline between antennas i and j, in a fringe-fit interval $\tau_{\rm acc}$ (see Lecture 9):

$$\Delta S_{ij} \;=\; \frac{\sqrt{T_{{\rm sys},i}\; T_{{\rm sys},j}}}{C\eta_s\sqrt{(2\;\Delta\nu\;\tau_{\rm acc}\;K_i\;K_j)}}\;. \tag{26–3}$$

Here, $K_i = (\eta_{a,i}A_i)/(2k_B)$, where $\eta_{a,i}$ is the aperture efficiency of antenna i, A_i is its physical area, and k_B is Boltzmann's constant. In addition $T_{{\rm sys},i}$ is the system temperature of antenna i; C is the coherence ($C \equiv 1$ for perfect coherence); η_s is the system efficiency factor (due to 1- or 2-bit sampling, fringe rotator quantization, etc.); and $\Delta\nu$ is the observing bandwidth.

Often, we express the sensitivity of a telescope in Jansky by dividing its system temperature (in K) by its gain (in K/Jy), to derive a quantity known as the "System-Equivalent Flux Density" (SEFD). Then,

$$\mathrm{SEFD}_i \;=\; \frac{T_{{\rm sys},i}}{K_i}\;, \tag{26–4}$$

and

$$\Delta S_{ij} \;=\; \frac{\sqrt{\mathrm{SEFD}_i\;\mathrm{SEFD}_j}}{C\eta_s\sqrt{2\;\Delta\nu\;\tau_{\rm acc}}}\;. \tag{26–5}$$

Fringe detection requires a signal-to-noise ratio of approximately 7 to minimize the probability of a false detection (see Thompson, Moran, & Swenson 1986, pp. 259–269); i.e., the minimum correlated flux density of the target source must be $S_{\rm c,min} \geq 7\Delta S_{ij}$. As an example, we compute $S_{\rm c,min}$ for two 25-m VLBA antennas at 5 GHz, assuming a 300-second integration over a 16-MHz bandpass, with 2-bit sampling. For these antennas, SEFD $\approx$ 300, and $\eta_s = 0.88$ for 2-bit sampling, so $S_{\rm c,min} \geq 24$ mJy. However, for SVLBI, the space telescope is usually much less sensitive than a ground telescope. In particular, the 5-GHz SEFD of HALCA is approximately 16,000 Jy, so $S_{\rm c,min} \geq 180$ mJy.

A variety of fringe-fitting techniques have been investigated in processing VSOP data. Accurate amplitude calibration is helpful, so that global fringe-fitting programs can give appropriate weight to the strongest baselines. Parameters in the AIPS task FRING enable different methods of combining or "stacking" the baselines in the fringe search, which can improve the detection threshold by critical factors of 1.5–2. Restricting the size of the fringe-search window is often important, since this limits the number of cells that must be searched, reducing the strength of the largest noise spike expected. Finally, it is sometimes useful to set a very low signal-to-noise threshold in order to accept all FRING solutions; the residual delays and rates can be plotted to search for repeating values, which may indicate the presence of weak fringes.

4.4. Coherence

It is desirable to use the longest possible integration time, τ_{acc}, in fringe fitting, as long as the coherence $C \approx 1$ (see Equation 26–5). For VSOP, coherence times of $\tau_c \approx$300–400 seconds are typical at 5 GHz, while τ_c may be 500–600 seconds at 1.6 GHz. (Here, τ_c is defined to be the interval over which C falls to 0.9.) These are in good agreement with the values predicted prior to launch, based on the expected orbit accuracy, propagation effects in the Earth's atmosphere, and the properties of the phase-transfer process. Coherence times at both frequencies are often reduced by factors of 2–4 within 30–60 minutes of perigee, apparently due to increased errors in the reconstructed velocity at low altitudes. We note the dependence of the coherence on the r.m.s. phase residual, $\Delta\phi_{\mathrm{rms}}$:

$$C = 1 - \frac{(\Delta\phi_{\mathrm{rms}})^2}{2} \ . \tag{26–6}$$

After a delay which is a linear function of time has been fitted and removed, if the time derivative of the residual fringe rate is defined as $\dot{\nu}_0$, the integration time for 90% coherence can be shown to be

$$\tau_c \approx 375 \left(\frac{\dot{\nu}_0}{10\ \mathrm{mHz\ hr^{-1}}} \right)^{-1/2} \ \mathrm{sec}\ . \tag{26–7}$$

Figure 26–8 shows the impact of the Time Corrections File on the fringe-fit residuals and coherence for a 5-GHz VSOP observation of 3C 345, on 21 March 1998. The top panels show the residual delay and fringe rate for the Space Radio Telescope in correlation of a 1.75-hr segment of the observation without use of a Time Corrections File, from 03:50 to 05:35 UT. (For reference, note that perigee was at 06:09 UT.) The fringes were centered at the initial time by extracting a constant clock offset from the Time Corrections File; otherwise no fringes could have been found. The residual fringe rate approaches the maximum value of ±3.8 Hz for the VLBA correlator, and the change in that rate ranges from $\sim$ 300 mHz hr^{-1} near 04:00 to $\sim$ 1100 mHz hr^{-1} near 05:30. Equation 26–7 then implies coherence times ranging from 35 to 70 sec. The bottom panel shows the results of a correlation of the same data set using the Time Corrections File, which was shown previously in Figure 26–5. Note that the ranges of residual delay and fringe rate are both reduced by more than two orders of magnitude, The maximum variation of the residual rate is $\sim$ 10 mHz hr^{-1}, implying $\tau_c \approx$ 375 sec. This increase in τ_c by a factor of $\sim$ 9 causes the noise to be reduced by a factor of $\sim$ 3, which can enable detection of SVLBI fringes for many more sources. The residual rate after application of the Time Corrections File, which corresponds to a velocity of about 1 mm sec^{-1}, is almost surely due to the effect of orbit errors on the interferometer model.

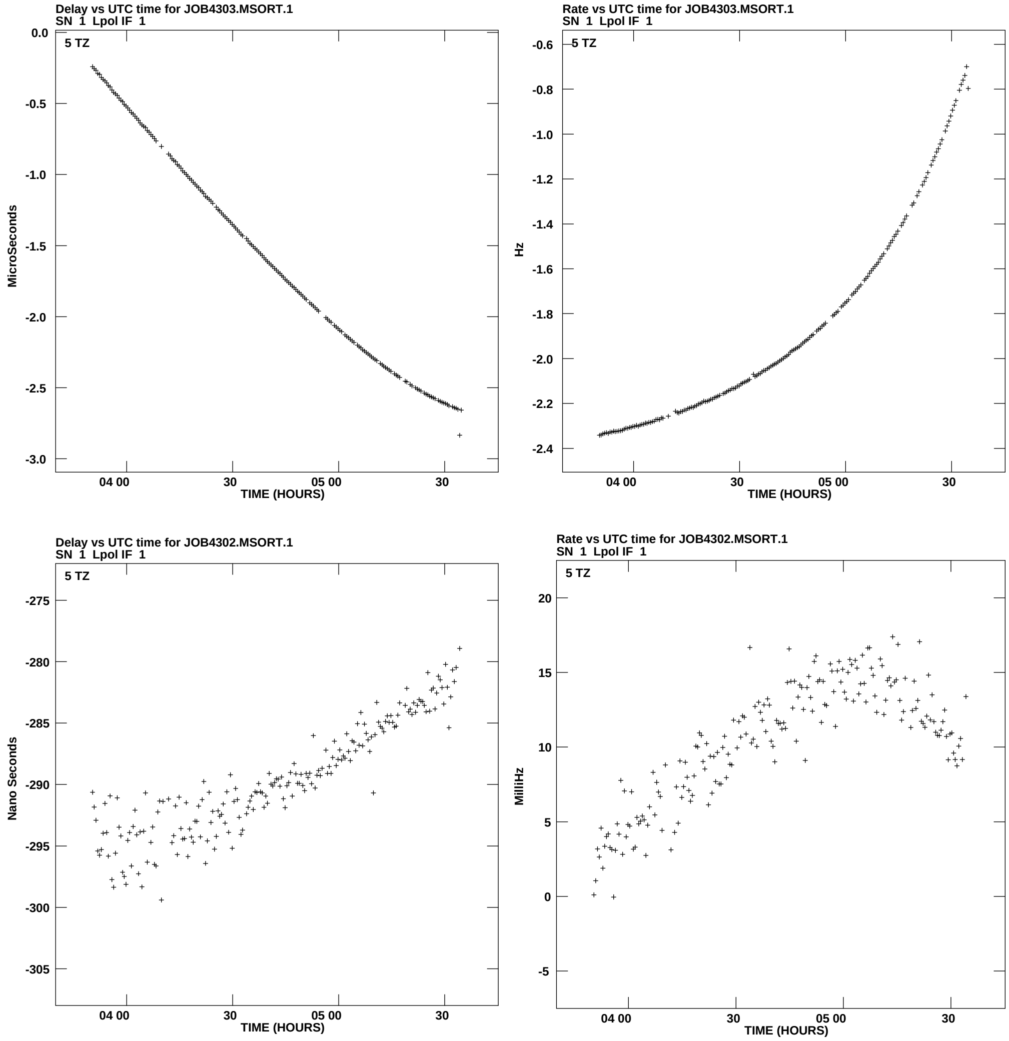


Figure 26–8. Residuals to fringe fits for a 5-GHz observation of 3C 345 using HALCA and the Tidbinbilla tracking station. Residual delays are on the left, and residual fringe rates on the right. *Top:* Results for correlation without a Time Corrections File. *Bottom:* Results for correlation using the Time Corrections File shown in Figure 26–5.

5. Imaging and Modeling

It has taken many years to establish imaging and model-fitting techniques for ground VLBI. Scientific data from VSOP have been available for only a year, and we are just learning the best procedures to use for SVLBI. Imaging is typically done using DIFMAP (Shepherd 1997) in the Caltech VLBI package, or either IMAGR or SCMAP in AIPS (van Moorsel, Kemball, & Greisen 1996). This section will be limited to somewhat general comments; major progress is expected over the next several years.

5.1. Data Weighting

Data weighting during fringe fitting depends on the sensitivity of the individual baselines, as derived from the *a priori* calibration. In imaging, one can use weighting ranging from purely "natural" (equal weight for each visibility point) to purely "uniform" (equal weight for each grid cell); Lecture 7 describes a "robustness" parameter that can be used to adjust the weights between the two extremes (see also Briggs 1995). It is also possible to "taper" the data, giving higher weight to the shorter projected baselines. In SVLBI, in order to achieve the highest resolution, it is desirable to give a high weight to the baselines involving the orbiting antenna. Therefore, natural weighting and significant tapering are generally undesirable, since both will down-weight the space baselines and make the final image approach that from a ground-only VLBI observation.

5.2. Self-Calibration

Self-calibration is a critical element of SVLBI. Sampling of the (u,v) plane is poorer than for the VLBA, particularly for southern sources observed with sparse ground arrays. Often, the sampling leads to a dirty beam with sidelobes as high as 30%–50% of the main beam, so calibration errors will scatter large amounts of flux around the map. The spacecraft typically is at the end of many similar long baselines, so the closure phase used in phase self-calibration is poorly constrained (Linfield 1986). Some observations include only 3 or 4 ground telescopes with relatively poor *a priori* calibration. Phase self-calibration (requiring groups of 3 telescopes) and amplitude self-calibration (requiring 4 telescopes) may not be very effective in improving images for these small ground arrays.

Careful data editing is more important for SVLBI than for the VLBA because of the difficulties with self-calibration. In particular, tracking stations may have delay jumps or multiple clock initializations. It is important to identify these events, and either flag the data or break the fringe fits at appropriate times. One technique used by the author is to perform amplitude self-calibration with a short solution interval (minutes), inspect the solutions for large corrections to the spacecraft gain, and use this to identify data that should be flagged before performing the "real" self-calibration.

An important aspect of self-calibration is to start off with some knowledge of the structure of the program source. Even simple information about the jet direction helps greatly; otherwise, large sidelobes may make it difficult to distinguish real structures from spurious components, and self-calibration will not easily remove false symmetries. Most of the strong sources observed by HALCA have been imaged at 2.3 and 8.5 GHz (Fey, Clegg, & Fomalont 1996), and the on-line images (*http://maia.usno.navy.mil/rorf/rrfid.html*) can be extremely valuable. Ground-based VLBA images at 15 GHz (Kellermann et al. 1998; see *http://www.cv.nrao.edu/2cmsurvey/*) also are quite useful.

5.3. Dynamic-Range Limitations

It is difficult to quantify the dynamic range possible for SVLBI observations. For HALCA observing with the VLBA, dynamic ranges (the ratio of the peak flux density to either the average map noise or the strongest spurious feature) of a few hundred have been achieved with moderately good sampling of the (u,v) plane (see Section 5.4.). On the other hand, for VSOP survey observations

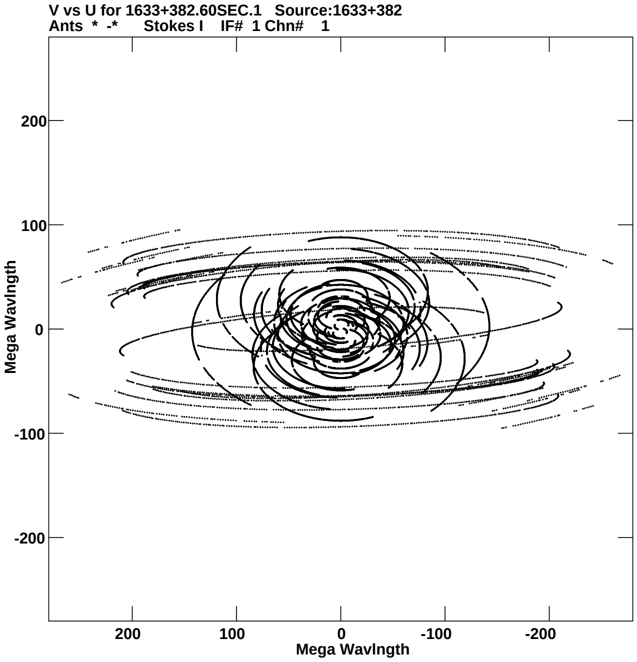


Figure 26–9. (u,v) coverage for HALCA+VLBA observations of 1633+382 on 29/30 July 1997.

involving only 2–3 ground telescopes, dynamic ranges of only 10–20 may be achievable. For these data, model-fitting in the (u,v) plane may be the best way of characterizing the source (see Lecture 160. Classical model-fitting techniques weight the data by the baseline sensitivities, but this gives low weight to the space baselines, hardly a desirable circumstance. Since the techniques of SVLBI model-fitting are still poorly explored, no further details will be given here.

5.4. Sample Image

A VSOP observation of the γ-ray blazar 1633+382 (Ulvestad et al. 1998) can be used to illustrate the results of SVLBI imaging. This source was observed at 5 GHz by HALCA and the VLBA during the VSOP in-orbit checkout, on 29/30 July 1997. The ground telescopes spent about 9 hr on source, while 5 hr of useful HALCA data were supplied via the tracking stations in West Virginia and California. The (u,v) coverage for this observation is shown in Figure 26–9; the projected baseline lengths were less than half the maximum achievable for VSOP. Therefore, the (u,v) plane was well filled, giving good imaging capability, but sub-optimal resolution. In fact, the North-South resolution was not much better than that of the VLBA, while the East-West resolution was a factor of ~ 2 better than for the ground-only array.

Two resulting images of 1633+382 are shown in Figure 26–10. The left-hand panel shows the ground-only image, while the right-hand panel shows the image made including HALCA. In this case, the improvement in the East-West resolution was critical for separating the central source into two distinct

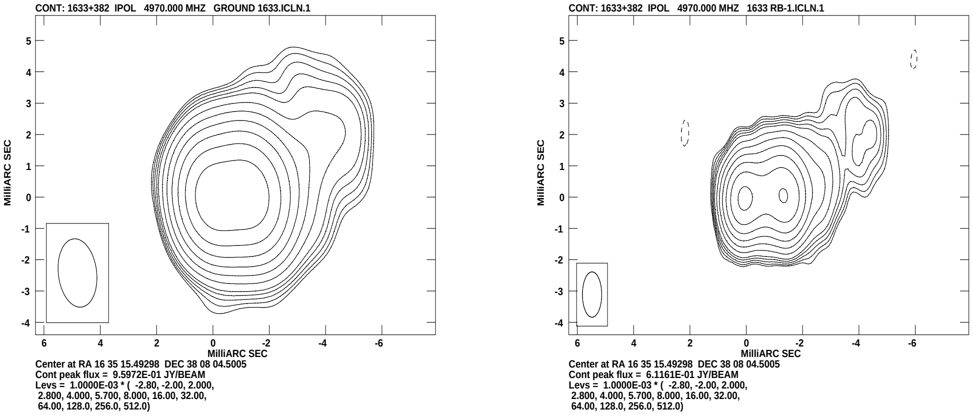


Figure 26–10. Two 5-GHz images of γ-ray blazar 1633+382, at identical scales, from data taken on 29/30 July 1997. *Left:* ground image, using only the VLBA telescopes. *Right:* SVLBI image, using the VLBA and HALCA.

components. The dynamic range of the VSOP image is several hundred to one.

6. Future Missions

6.1. RadioAstron

RadioAstron (Kardashev 1997) is a first-generation SVLBI mission being developed under the leadership of the Astro Space Center in Moscow. The RadioAstron launch was scheduled to precede HALCA's launch, but financial problems have delayed it until after the turn of the millennium. RadioAstron will carry a 10-m radio telescope that is planned to operate in four frequency bands: 0.3, 1.6, 4.8, and 22.2 GHz. The nominal orbit is a 28-hr orbit with an apogee height of nearly 80,000 km and an initial perigee height of 4,000 km, although there also has been talk of a 4-day or even an 8-day orbit. RadioAstron will use a set of tracking stations similar to those available for VSOP, but with a station in Russia replacing the station in Japan. The link frequency for the tracking stations will be 8 GHz rather than the 15-GHz frequency used for HALCA, so ionospheric effects on the link may reduce the coherence slightly.

6.2. VSOP-2

VSOP-2 is a possible successor to VSOP. A formal working group has been formed at ISAS, and a proposal for a new start for VSOP-2 is expected in 1999 or 2000. If the mission is approved, the launch date would be about 2006 or later. The VSOP-2 spacecraft would have an orbit similar to VSOP, with a 10-m antenna, cooled receivers, frequency coverage up to 43 GHz, and a data rate of about 1 Gbit sec^{-1}. Therefore, the interferometer sensitivity of VSOP-2 would be about 10 times better than VSOP.

6.3. ARISE

ARISE (Advanced Radio Interferometry between Space and Earth) is a mission concept currently being studied within NASA's Structure and Evolution of the Universe theme. It is in the long-term "roadmap" for that theme, with a possible launch envisioned in 2008 or later. ARISE is an ambitious mission concept, whose basic goal is to launch a VLBA-equivalent telescope into an elliptical orbit with an apogee height of roughly 40,000 km. The telescope would be a 25-m inflatable antenna with cooled receivers, frequency coverage as high as 86 GHz, and a data rate of 1–8 Gbit $\sec^{-1}$. Further details on this mission concept can be found in a number of references (e.g., Gurvits, Ulvestad, & Linfield 1996; Ulvestad, Gurvits, & Linfield 1997; Ulvestad & Linfield 1998).

Acknowledgments. I thank the many participants in the international Space VLBI endeavors for their efforts, David Murphy for development and distribution of the FAKESAT software, and Mark Claussen and Phil Edwards for reviewing earlier drafts of this lecture. Special gratitude is due to Gerry Levy, who pioneered Space VLBI by making the TDRSS experiments happen, and to Roger Linfield, for years of collaboration and discussions on all aspects of Space VLBI. Space VLBI efforts at NRAO are funded by the National Aeronautics and Space Administration.

References

Bate, R. R., Mueller, D. D., & White, J. E. 1971, *Fundamentals of Astrodynamics,* (New York: Dover).

Beasley, A. J., & Conway, J. E. 1995, in *Very Long Baseline Interferometry and the VLBA*, ASP Conf Series 82, eds. J. A. Zensus, P. J. Diamond, & P. J. Napier (San Francisco: ASP), 327–343.

Benson, J. M. 1995, in *Very Long Baseline Interferometry and the VLBA*, ASP Conf Series 82, eds. J. A. Zensus, P. J. Diamond, & P. J. Napier (San Francisco: ASP), 117–131.

Boden, D. G. 1992, in *Space Mission Analysis and Design,* 2nd edition, eds. W. J. Larson & J. R. Wertz (Torrance, CA: Microcosm, Inc.), Chapter 6.

Briggs, D. 1995, *High Fidelity Deconvolution of Moderately Resolved Radio Sources,* Ph.D. thesis, New Mexico Institute of Mining and Technology.

D'Addario, L. 1991, *IEEE Trans. on Instrumentation and Measurement,* 40, 584–590.

Fey, A. L., Clegg, A. W., & Fomalont, E. B. 1996, *ApJS*, 105, 299–330.

Griffin, M. D., & French, J. R. 1991, *Space Vehicle Design* (Washington: American Institute of Aeronautics and Astronautics), Chapter 4.

Gurvits, L. I., Ulvestad, J. S., & Linfield, R. P. 1996, in *Large Antennas in Radio Astronomy,* ed. C. G. M. van't Klooster (Noordwijk: ESTEC), 81–88.

Hirabayashi, H. 1998, in *IAU Colloquium 164: Radio Emission from Galactic and Extragalactic Compact Sources*, ASP Conf Series 144, eds. J. A. Zensus, G. B. Taylor, & J. M. Wrobel (San Francisco: ASP), 11–15.

Hirabayashi, H., et al. 1998, *Science*, 281, 1825–1829.

Hirosawa, H., & Hirabayashi, H. 1995, in *IEEE AES Systems Magazine,* June 1995, 17–23.

Kardashev, N. S. 1997, *Experimental Astronomy,* 7, 329.

Kawaguchi, N., Kobayashi, H., Miyaji, T., Mikoshiba, H., Tojo, A., Yamamoto, Z., & Hirosawa, H. 1994, in *VLBI Technology: Progress and Future Possibilities,* eds. T. Sasao, S. Manabe, O. Kameya, & M. Inoue (Tokyo: Terra Scientific), 26–33.

Kellermann, K. I., Vermeulen, R. C., Zensus, J. A., & Cohen, M. H. 1998, *AJ*, 115, 1295–1318.

Kemball, A., et al. 1998, *Polarization VLBI observations at 1.6 GHz and 5 GHz with the HALCA satellite: first results from in-orbit checkout observations,* available from *http://www.vsop.isas.ac.jp/obs/Pol.html.*

Levy, G. S., et al. 1986, *Science,* 234, 187–189.

Levy, G. S., et al. 1989, *ApJ*, 335, 1098–1104.

Linfield, R. P. 1986, *AJ*, 92, 213–218.

Linfield, R. P., et al. 1989, *ApJ*, 335, 1105–1112.

Linfield, R. P., et al. 1990, *ApJ*, 358, 350–358.

Meier, D. L. 1998, in *IAU Colloquium 164: Radio Emission from Galactic and Extragalactic Compact Sources*, ASP Conf Series 144, eds. J. A. Zensus, G. B. Taylor, & J. M. Wrobel (San Francisco: ASP), 421–422.

Murphy, D. W., et al. 1994, in *VLBI Technology: Progress and Future Possibilities,* eds. T. Sasao, S. Manabe, O. Kameya, & M. Inoue (Tokyo: Terra Scientific), 34–38.

Murphy, D. W. 1995, *BAAS,* 186, 2706

Natori, M. C., Takano, T., Miyoshi, K., Inoue, T., & Kitamura, T. 1994, in *VLBI Technology: Progress and Future Possibilities,* eds. T. Sasao, S. Manabe, O. Kameya, & M. Inoue (Tokyo: Terra Scientific), 10–20.

Pilbratt, G. 1991, in *IAU Colloquium 131: Radio Interferometry: Theory, Techniques and Applications,* ASP. Conf. Series 19, eds. T. J. Cornwell & R. A. Perley (San Francisco: ASP), 102–106.

Romney, J. D. 1995 in *Very Long Baseline Interferometry and the VLBA*, ASP Conf Series 82, eds. J. A. Zensus, P. J. Diamond, & P. J. Napier (San Francisco: ASP), 17–37.

Schilizzi, R. T., et al. 1984, in *IAU Symposium 110: VLBI and Compact Radio Sources,* eds. R. Fanti, K. Kellermann, & G. Setti (Dordrecht: Reidel), 407–414.

Shepherd, M. C. 1997, in *Astronomical Data Analysis Software and Systems VI,* ASP Conf. Series 125, eds. G. Hunt & H. E. Payne (San Francisco: ASP), 77–84.

Takano, T., Natori, M., Ohnishi, A., Miura, K., Inoue, T., Noguchi, T., & Kitamura, T. 1994, in *Proceedings of the 19th International Symposium on Space Technology and Science,* eds. M. Hinada, Y. Arakawa, Y. Horikawa, J. Kawaguchi, T. Nakajima, & I. Nakatani (Tokyo: Agne Shofu), 483–492.

Thompson, A. R., Moran, J. W., & Swenson, G. W. 1986, *Interferometry and Synthesis in Radio Astronomy* (New York: Wiley & Sons).

Ulvestad, J. S., Gurvits, L. I., & Linfield, R. P. 1997, in *High Sensitivity Radio Astronomy,* eds. N. Jackson & R. Davis (Cambridge: Cambridge University Press), 252–255.

Ulvestad, J. S., & Linfield, R. P. 1998, in *IAU Colloquium 164: Radio Emission from Galactic and Extragalactic Compact Sources*, ASP Conf Series 144, eds. J. A. Zensus, G. B. Taylor, & J. M. Wrobel (San Francisco: ASP), 397–398.

Ulvestad, J. S., Vestrand, W. T., Stacy, J. G., & Biretta, J. A. 1998, in preparation.

van Moorsel, G., Kemball, A., & Greisen, E. 1996, in *Astronomical Data Analysis Software and Systems V,* ASP Conf. Series 101, eds. G. H. Jacoby & J. Barnes (San Francisco: ASP), 37–43.

VSOP Science Operations Group 1998, *AO2 Proposers Guide, VLBI Space Observatory Program,* ed. D. W. Murphy, available from http://www.vsop.isas.ac.jp.

Walker, R. C. 1984, in *Indirect Imaging,* ed. J. A. Roberts (Cambridge: Cambridge University Press), 53–65.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

27. Interferometric Array Design

M.A. Holdaway & Tamara T. Helfer
National Radio Astronomy Observatory, Tucson, AZ 85721, U.S.A.

Abstract. We investigate some of the principles which lead to the design of radio interferometric arrays and array configurations, including both abstract issues such as sensitivity and Fourier plane coverage, and practical issues such as moving antennas and site topographical constraints. We draw on the design and history of existing arrays and also give a glimpse of what ideas and algorithms are helping design new instruments such as the Submillimeter Array (SMA) and the Millimeter Array (MMA).

1. Introduction

Array design can include a variety of topics: how many antennas should the telescope have, and how big should they be? Are there astronomical requirements which dictate an aspect of the array layout? How many antenna configurations will there be, and how will the different configurations work together? How should we design each individual configuration? But the central topic of array design deals with how to efficiently sample the Fourier plane. Each interferometer, or pair of antennas, at a given moment in time samples a single point in the Fourier plane, and we need to arrange the antennas in such a way that the set of sampled points permits us to make high quality, high sensitivity images. Since most antennas require a fair amount of infrastructure with not insubstantial capital costs on the ground beneath their bases (called *antenna pads*), it is important to design a good set of antenna configurations which adequately sample the Fourier plane before the array is built.

This chapter will be of interest to anyone who is building an interferometric array in their back yard. Beyond this, it will also help people who are using modern synthesis telescopes understand why their telescope looks the way it does. This information might even influence the way they use the telescope or the way they process their data. We don't deal with historical telescopes that don't operate anymore. But that doesn't stop us from examining telescopes which don't yet exist.

2. Telescope Design

Long before we think about how to lay out our antennas, we need to consider what we are going to lay out: how many antennas n will we build, and how big will they be (diameter D)? The answer to these questions is tied up in the science that we will do with the array and how much money we will have to build it. Is it wise or dangerous to guess what science will be done with an array before it is built? The bulk of the VLA's current observations were never envisioned by the designers of the VLA, but because they built a flexible instrument which did not preclude these observations, the VLA has been able to do great and new things anyway. So, we must be so clever that we are not stung by our cleverness.

Point source sensitivity: optimizing nD^2. If we are interested in raw sensitivity for observations of unresolved sources, or sources which will require only a single pointing on the sky, then the equations of Lecture 9 indicate that we should maximize the collecting area of the array, nD^2. To understand what that means for the design of an array, we need to know how much the electronics for each antenna will cost, how the cost of the correlator will vary with the number of antennas, and how the cost of the antennas scales with D. We can achieve the same amount of collecting area with a very large number of tiny antennas, a fair number of regular sized antennas, or a few very large antennas. Building a very large number of tiny antennas is generally expensive because the costs of the per antenna electronics and the correlator eventually dominate the cost of the array. Building a few very large antennas will save a lot of money on the per antenna electronics costs and correlator costs, but building very large antennas which meet the required specification becomes prohibitively expensive. So, the cheapest way to build an array with a given collecting area is a compromise between the many small and the few large antennas. Optimizing for nD^2 will bias the telescope design towards fewer, slightly larger antennas.

Imaging sensitivity: optimizing nD. There are two different measures which indicate that nD should be optimized, and they both relate to imaging. First, consider that we are interested in making images of extended sources which are larger than the primary beam, and hence require multiple pointings to image. Most VLA observations require only a single pointing, but the large VLA sky surveys are good examples of multiple pointing observations. Because of the very short observing wavelengths, and hence the very small primary beams, the MMA will require multiple pointings for imaging a large fraction of the time. The primary beam size is given by λ/D, so the number of pointings required to cover a given region of the sky will be proportional to D^2, or the time spent integrating on each pointing will be proportional to D^{-2}. Sensitivity, the inverse of noise, is proportional to $\sqrt{t}$, and hence there will be a factor of D^{-1} in sensitivity due to the number of pointings required to cover a given region of the sky. This needs to multiply the nD^2 from the point source sensitivity, and we find that imaging sensitivity is proportional to nD. If we design an interferometric telescope which optimizes nD, we would be biased towards more, slightly smaller antennas.

The second array criterion related to imaging is the Fourier plane coverage. For our current purposes, we will reduce the entire complicated issue of Fourier plane coverage into a single number, the fraction of filled cells (Hjellming 1989). In the image plane, we'll image the entire primary beam of size λ/D. In the Fourier plane, that means our cell size will be of order D/λ. Assume that we are comparing among interferometric telescopes of the same maximum baseline (i.e., resolution), so that the extent of the Fourier plane will be b_{max}/λ. Then the number of cells in the Fourier plane that we need to fill is about $(b_{\mathrm{max}}/D)^2$. We fill those cells with the $n(n-1)/2$ visibilities that we collect every integration time. Assuming that these visibilities are non-redundant (i.e., each one goes into its own cell), then the fraction of filled cells will be proportional to $n(n-1)D^2/(2b^2_{\mathrm{max}})$. Since b_{max} is assumed to be the same for arrays being compared, the criterion is $(nD)^2$. Hence, by optimizing nD for a given telescope budget, we also optimize both the wide field (i.e., multiple pointing) imaging sensitivity and the fraction of occupied cells, which is a measure of the Fourier plane coverage quality.

Scientific Drivers: A few examples of how science interacts with the telescope and array design follow.

Nobody ever does blind surveys on the sky with VLBI. Nobody ever mosaics with VLBI. VLBI people are interested in point source sensitivity and Fourier plane coverage. Point source sensitivity is desired to detect faint sources in the atmospheric coherence time. Traditionally, ad hoc VLBI arrays have been starved for Fourier plane coverage, and the complexity of source structure seen in VLBI images continues to evolve as the VLB arrays improve. While there is no generally accepted statement of how many antennas gives good enough Fourier plane coverage for a VLBI experiment, it is clear that VLBI pushes us towards large dish sizes, ie. ≥ 25 m.

The Nobeyama Radioheliograph was designed to image the sun at 17 and 34 GHz. In order to fit the sun into its primary beam at the high frequency, very small (80 cm) dishes were built. With such small dishes and uncooled receivers, they could afford to build very many dishes, which is required to make snapshot images of the sun. Snapshot imaging is required since the radio emission from the active sun can change on sub-second time scales.

The MMA was originally conceived as an array of 40 8 m dishes which would spend a large fraction of its time observing objects which required multiple pointings. With 40 8 m dishes, the MMA appeared to optimize nD for the amount of money NRAO was requesting.

The Large Southern Array (LSA), proposed by ESO, is also a large millimeter wavelength interferometer, but the European scientific community was mainly interested in young galaxies at high z, which are found to be compact. For this reason, the LSA was conceived as 50 16 m antennas, which was claimed to optimize nD^2, the point source sensitivity.

Currently, NRAO and ESO have begun a process in which the MMA and the LSA will be combined into a single joint array. Either the two design groups will have to compromise on the antenna diameter, or the joint array will be a heterogeneous array.

Homogeneous and Heterogeneous Arrays. A homogeneous array is an array in which all antennas are the same size. A heterogeneous array will have antennas of two or more diameters. It is not problematic for VLBI arrays to be heterogeneous because they are almost always observing sources which are very small compared to the primary beam. However, homogeneity becomes an issue when we observe sources which are comparable to or larger than the primary beam.

Cornwell, Holdaway, & Uson (1993) demonstrated that sources much larger than the primary beam could be accurately imaged using homogeneous array mosaicing, i.e., by measuring interferometric data and total power data with antennas of the same size. Hence, a heterogeneous array is not required to measure short baseline information. However, Holdaway (1997a, 1997b) and Wright (1997) have also demonstrated that the mosaicing algorithms work fine *with* a heterogeneous array. There may be logical or political reasons why one would want to design and build a heterogeneous array (for example, the combining of two existing arrays which have different dish diameters).

Mosaicing and Field of View as Design Drivers. Sometimes, the way images are made by an instrument will influence the design of that instrument.

For the VLA, we did not know a great deal about how to make images when it was designed, but it worked out pretty well anyway. The MMA will spend a large fraction of its time, perhaps as much as or more than 50%, imaging objects that are larger than the antennas' primary beams. In this aspect, the MMA is fundamentally different from any radio interferometer ever built, and the mosaicing requirement results in many specifications which are tighter than normal interferometers. The tightened pointing and surface accuracy specifications are mentioned in Lecture 20.

If an interferometer is observing sources much smaller than the primary beam, there is no particular need to completely fill the Fourier plane in an absolute sense. For a source size $\theta_{\max}$, the size of the cells in the (u, v) plane will be like $1/\theta_{\max}$. Hence, for a constant maximum baseline, observing a smaller source means larger (u, v) cells and hence fewer (u, v) cells. For a well designed array, this means that more of the (u, v) cells will have data in them, or that the Fourier plane coverage is more nearly complete, at least for this observation. Hence, the typical target source size will impact the required Fourier plane coverage and the number of antennas.

For mosaicing, the source fills the beam, so we really *do* need something like complete Nyquist sampled Fourier plane coverage. As we learned in Lecture 20, the effective Fourier plane coverage for mosaicing is the nominal δ function (u, v) coverage convolved with the autocorrelation of the antenna illumination. It is the effective mosaicing Fourier plane coverage which should be essentially complete. The MMA design will have a $\sim$90 m configuration which will provide complete effective snapshot coverage, which is highly desirable for very large mosaics, and a $\sim$250 m configuration which will provide nearly complete effective coverage in a few snapshots.

3. Multi-configuration View

Most existing interferometers are reconfigurable. Understanding how the different configurations fit together is an integral part of the design of an interferometric array.

3.1. Why multiple Configurations?

The DRAO Synthesis Telescope in Penticton has seven antennas, four fixed and three movable, all on an E-W line with a 600 m maximum baseline. Essentially complete Fourier coverage out to the maximum baseline is obtained by observing 12 hour tracks in each of twelve different configurations. In this case, the different configurations are required to obtain good Fourier plane coverage.

The VLA, on the other hand, has 27 antennas and gets pretty good, but not complete, Fourier plane coverage in each of its four scaled configurations. For the VLA, good images can often be obtained from a single configuration observation, and the different configurations are provided mainly to give different resolutions or brightness sensitivities. Faint, extended sources can be detected in the compact D array where the beam is largest and the brightness sensitivity is the best. Bright sources with compact features can be further studied in detail at high resolution with larger arrays. And since the resolution in radians is about $\lambda/b_{\max}$, bright objects can be studied at the same resolution at multiple

λ by picking the right configuration for the observing wavelength. The scaling of the VLA configurations (3.285) fits approximately with the scaling between the VLA's original observing bands at 20, 6, and 2 cm. Since the VLA's four configurations are scaled versions of each other, the large hole in the center of the Fourier plane grows larger and larger with each array until the A array has a ~1 km hole at its center. So, another function of the VLA's multiple configurations is to provide short spacings when imaging large objects at high resolution.

A notable radio astronomer once asked if the MMA was going to finally get away from the nonsense of having multiple configurations. After all, with something like 40 antennas, won't the Fourier plane coverage be good enough to get away with a single configuration? This person was confusing the way Penticton uses multiple configurations (out of necessity due to lack of antennas) with the way a well designed, sufficiently funded interferometric array uses multiple configurations (to increase its flexibility). If a single configuration were used for the MMA, observations which required low resolution (or high brightness sensitivity) would have to taper their data (i.e., throw away the long baselines). Consider first a single configuration with a uniform Fourier plane distribution. If a 3000 m array were built and a resolution corresponding to a 300 m array were required, we would be left with only about 1/100 the baselines, or about 1/10 the sensitivity. So, while the brightness sensitivity would increase because the beam size increased more than the point source sensitivity decreased, we could gain by about a factor of 10 in sensitivity if we just built a 300 m array which did not require any tapering. An alternative single configuration which would not lose quite as much sensitivity to tapering would be one which produced a highly condensed Fourier plane coverage. However, in this case, we would need to perform uniform weighting (i.e., down weighting the oversampled short spacings) in order to achieve the high resolution imaging, at a loss in sensitivity. So, to optimize the sensitivity of an interferometric telescope, it makes sense to have multiple configurations.

3.2. But How Many Configurations?

If we follow this reasoning to its logical end, we will have a different configuration for each observation, which would cost too much money. For the MMA, we have estimated the optimal number of configurations through a cost-benefit analysis (Holdaway 1998a). In this analysis, we trade off the expenses associated with multiple configurations (i.e., antenna transporters, additional antenna stations, lost observing time due to reconfiguration and recalibration, and workers' salaries to move the antennas) against the benefits of having high sensitivity observations due to not having to taper very much because there is a configuration which is about right for any particular observation. Unlike the VLA, there are no favored frequency bands that can be used to argue for configuration scaling. Rather, the MMA will mainly observe spectral lines which are found at all the frequencies in the range from 30 GHz to 950 GHz (except at the atmospheric absorption lines). Multi-transition studies comparing one species to another at the same resolution will be common, and we will have to taper some of the data to get to a common resolution. For the assumed distribution of desired resolutions, including a fraction of observers who did not care about resolution, we concluded that about four or five configurations is optimal for the MMA. These results depend upon assumptions such as how many antennas must be moved between different configurations. With pure ring arrays, there are not very many pads which can be shared among the different arrays and almost all of the antennas must be moved when the configuration is changed. However, Conway (1998) and Webster (1998b) have proposed centrally condensed configuration schemes which can be gradually and continuously reconfigured by moving just a few antennas from inner to outer or vise versa. The advantage of such a scheme is that the desired resolution can be achieved without down weighting any data, thereby achieving maximum sensitivity. The disadvantage is that the Fourier plane coverage is highly centrally condensed and the highest possible resolution for an array of some maximum baseline is not achieved.

3.3. Big Arrays Don't Need Big Central Holes

There is one aspect of the VLA's multiple configuration scheme which is quite non-optimal, and that is the very large central hole in the (u, v) coverage which requires multiple configurations for observing large sources at high resolution. Braun (1993) argues that since this central hole is actually a very small fraction of the Fourier plane being sampled by a given configuration, it should be fairly painless to better sample the Fourier plane within the central hole. He proposed variants for the VLA A, B, and C arrays which moved one or two antennas to provide more short baselines with the hope that these would often permit single configuration imaging at the VLA. The idea was further studied in the context of an altered C configuration with the shortest possible baselines by Holdaway (1994), and Rupen (1997, 1998) made observational tests of such a configuration. The proposal pressure overwhelmingly favors replacing the C configuration with the altered C configuration. Shortened versions of the A and B configurations are not currently planned, but should be. It is the MMA's goal to include very short baselines in every array to permit single configuration imaging much of the time.

3.4. Strategy to Combine Data From Multiple Configurations

In spite of the presence of very short baselines in each configuration, there will be times when multi-configuration data will be combined for the MMA. The VLA requires it more often, even after the central holes in the large configurations are fixed, as the fixes are not bullet proof. A general guideline for combining multi-configuration data is that we seek a "coherent" (u, v) distribution in the final concatenated data set. We don't want large discontinuities in the Fourier plane density as we traverse the (u, v) plane radially. These large discontinuities will result in bad beams, or fixing them will result in wasted sensitivity through down weighting the problems.

Each VLA configuration has a highly condensed (u, v) distribution already (see below). Adding multiple configurations to each other will make the concatenated data even more centrally condensed. As each configuration is scaled by the factor of 3.285 from its neighbor, the $n + 1$ configuration will have (u, v) coverage which is sparser by $3.285^2 = 10.8$ than the n configuration (increasing n means larger arrays). Hence, just to get multi-configuration data which isn't more centrally condensed than single configuration data already is, we need to spend something like 11 times more time integrating in the B array than in the C array, or 121 times more time in A array than in C array.

The MMA doesn't yet have a strategy for combing multiple configurations. However, we do hope that all configurations will have enough short baselines to permit sufficient single configuration imaging most of the time. The MMA configurations won't produce (u, v) coverage as centrally condensed as the VLA configurations, and could be close to uniform. So, we can imagine shaping a beam through adding multi-configuration data together like a wedding cake. There may be a few edges that you need to smooth out via re-weighting the data.

4. Optimizing a Single Configuration: Pure Geometric Simplicity

Here, we deal with the real tofu of the topic of array configurations: how to arrange the antennas so that they give the Fourier plane coverage we want. It is worth thinking about the analogy with single dishes here. A single dish will have a resolution determined by its size and the observing wavelength, approximately $1.22\lambda/D$. A single dish, or filled aperture, can be thought of in terms of interferometry. Consider the dish to be made of lots of tiny little dishes touching each other. Instead of delaying and correlating the signals from each pair of tiny dishes, we reflect the signals to be combined in an analog fashion, by physically adding the electric fields together at the feed. In this picture of a single dish as being made up of lots of tiny dishes doing interferometry, we find that we have many short baselines and few long baselines. This is the natural distribution of Fourier samples for a filled aperture: the (u, v) density decreases monotonically with (u, v) distance until we reach the maximum baseline, which is the dish diameter. The details of this (u, v) distribution depend upon the details of the illumination pattern on the dish. This (u, v) distribution is inherently highly redundant.

We build interferometers because we greedy astronomers want high resolution, but we don't want to pay the cost of building outrageously large single dishes. We pay two prices for being cheap: an interferometer does not have the same brightness sensitivity as a single dish, losing by the array's filling factor; and we lose redundancy often to the point where a large fraction of the Fourier plane has NO samples. However, for an array, we do get some control over exactly where we want to put those Fourier samples. (There are also advantages of using an interferometer over a single dish. For example, a single dish measuring total power has a lot of signal power that must be removed before we get down to the astronomical signal. If these extraneous signals are not correctly removed, our noise level goes up. If the extraneous signals are incorrectly removed the same way all the time, we are left with systematic errors which will prevent the noise from averaging down. By correlating the detected signals from two dishes, the extraneous signals are usually not correlated, so interferometers are largely free from systematic errors.)

4.1. Abstract Fourier Plane Distributions

It is possible to study purely abstract Fourier plane distributions divorced from the reality of an antenna configuration. Instead of computing the (u, v) coordinates from some antenna array, the (u, v) coordinates can be generated from a probability distribution (Holdaway 1996). We mention three interesting cases: a Gaussian distribution, a distribution which increases with radial distance, and a uniform distribution, Each distribution is azimuthally symmetric, and the uniform and increasing distributions need to have some form of cutoff at the maximum baseline. The Gaussian (u, v) distribution is interesting because it will produce a Gaussian PSF (just think: no cleaning required). The radially increasing distribution is appealing because it will have the most baselines (and hence the most sensitivity) on the long baselines, right where the signal is the weakest for resolved sources. However, it is geometrically impossible to produce a 2-D array that gives you more long baselines than short, at least for arrays with more than three antennas. The best you can do on the long baselines is to get almost uniform Fourier plane coverage. However, uniform Fourier plane coverage with a sharp cutoff at the maximum baseline will result in a PSF which has very large inner sidelobes, like a J_0 Bessel function (i.e., the 2-D analog of a sinc function as the Fourier transform of a rectangular region).

Holdaway's (1996) studies indicate that a naturally tapered abstract (u, v) coverage produces better images than a uniform abstract (u, v) coverage. However, when simulations were performed with a randomly filled array, which produces a naturally tapered (u, v) coverage, and with ring-like arrays (i.e., a circle or a Reuleaux triangle), which have approximately uniform Fourier plane coverage, it was found that the ring-like arrays have superior imaging characteristics. Morita (1998) has demonstrated that the superior imaging of actual ring-like arrays is not due to the nearly uniform Fourier plane coverage, but instead is due to the excess of very short baselines in a ring array. If the antennas are arranged in a ring, the shortest baseline each antenna experiences will be shorter, on average, than shortest baselines experienced by the antennas in a filled (i.e., a fully 2-D) array. This emphasizes the importance of sufficient short baseline coverage for high quality imaging, and also the difference between abstract goals for the (u, v) coverage and the actual realities of arrays of antennas and their (u, v) coverages.

4.2. Linear Arrays

Linear arrays are nice because they are simple in many ways. It is easy to move the antennas about. (Often, linear arrays are on a railroad track.) Most linear arrays are oriented very nearly E-W; ideally, they are perfectly E-W. The snapshot coverage of a linear array is not very exciting, but for a source close to the celestial pole, each baseline of the E-W linear array makes a nice little circle in the Fourier plane. Figure 27–1 shows 1 hour tracks and the associated beam using the Westerbork Synthesis Radio Telescope. Because the (u, v) coverage is one dimensional, the beam is also one dimensional, perpendicular to the direction of the coverage. Figure 27–2 shows the (u, v) coverage and beam for a full 12 hour integration with the same linear array.

We need to make a little note about our figures of the antenna positions, the (u, v) coverage, and the synthesized beam. Normally, both the beam and the (u, v) coverage is viewed from the earth looking up. This makes sense, as the two will have a Fourier transform relationship and you can understand the pair intuitively. However, antenna locations are usually viewed from the sky

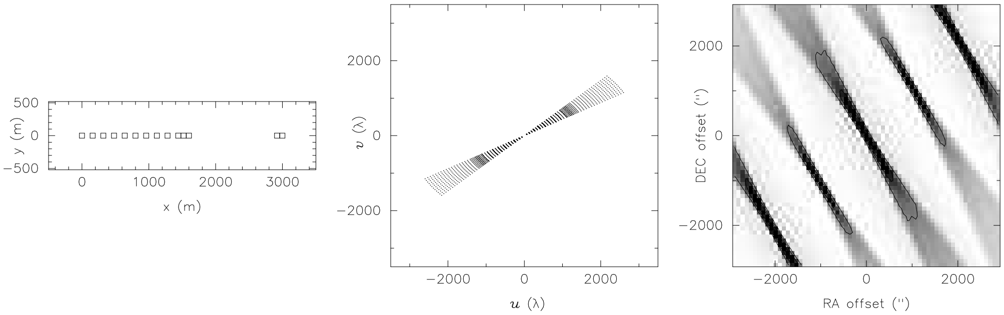


Figure 27–1. (Left) The antenna layout, (middle) the (u, v) coverage, and (right) the resulting synthesized beam for a 1 hour integration with the Westerbork linear array. The coverage is essentially one dimensional, and there will be no information about the source in the direction perpendicular to this coverage. The sign conventions of all three panels are such that the observer is looking down on the antennas, looking down on the (u, v) plane, and looking down on the synthesized beam. All centimeter wavelength observatories used as examples will have the (u, v) coordinates in wavelengths at 1 m wavelength. Millimeter wavelength examples such as the MMA will have (u, v) coordinates calculated at 1 mm wavelength.

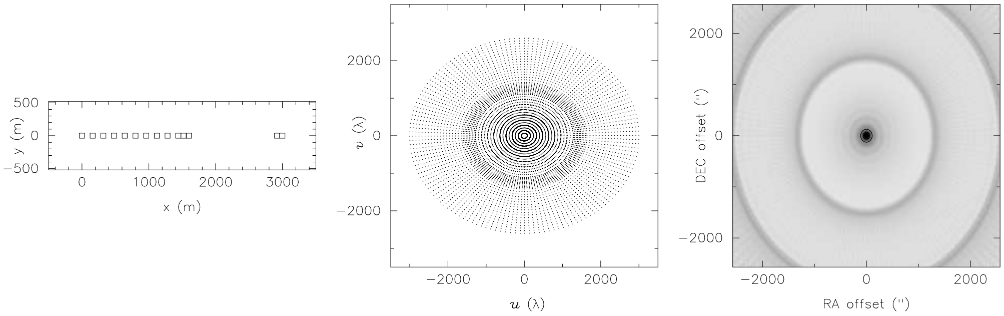


Figure 27–2. A 12 hour integration with the Westerbork linear array. Because the tracks are regularly spaced, large grating responses are found in the outer PSF.

looking down at the earth. Since there is a simple and intuitive connection between the antenna locations and the (u, v) spacings, we'd rather not make the reader flip the (u, v) coverage plots to relate them to the antenna positions. In fact, we'd like all three plots in each panel to be consistent, and we adopt the convention that we are *always* looking down from *outside* the celestial sphere, down onto the antennas on earth, down onto the (u, v) coverage projected on a plane perpendicular to the line of sight from our extra-celestial vantage point, and down onto the beam tangent to the celestial sphere. The latter two images are then backwards from what is normally displayed in order to appear consistent with the antenna locations. Furthermore, centimeter wavelength arrays such as the VLA, WSRT, GMRT, and VLBA will show (u, v) coverage and beams for a 1 m observing wavelength, while the MMA examples will show the coverage and beams at 1 mm wavelengths.

The Westerbork Synthesis Radio Telescope and the DRAO Synthesis Telescope configurations are both designed to have the distance between adjacent tracks in the Fourier plane be the same. Such a regular sampling in the Fourier plane results in a "grating response". Consider that the (u,v) coverage in a 1-D slice of the Fourier plane is a grid which is uniformly sampled at $\Delta b/\lambda$, multiplied by a rectangular region which is 1.0 inside the shortest baseline and 0.0 outside the longest baseline. Then the beam, given by the Fourier transform of the (u,v) coverage, will be a uniform grid in the image plane, separated by $\lambda/\Delta b$, convolved with a sinc function. In 2-D, instead of a grid, we have several concentric rings, and the Fourier transform will be a series of rings, or grating responses, separated by $\lambda/\Delta b$. If $\Delta b = D/2$, the Fourier plane is Nyquist (i.e., completely) sampled except for the central hole (you can't get baselines as short as $D/2$ or D), and the first grating ring will be located at $2\lambda/D$, which means that any source within the primary beam will make grating rings outside of the primary beam. Pretty good, eh?

Two disadvantages of E-W arrays are that snapshot imaging is impossible and the beam becomes increasingly elongated for sources near the equator. Two advantages of E-W arrays include the potential for complete Fourier plane coverage and simple low frequency imaging since all baselines in the array at all times will be coplanar.

Minimum Redundancy Linear Arrays. The resolution of a linear array which includes uniform sampling in the (u,v) plane may be maximized by minimizing the number of redundant spacings in the array (Moffet 1968). A completely nonredundant array is possible only up to 4 elements (Leech 1956): the spacings for a 4-element array are $*1*3*2*$, and this array includes all multiples of the unit spacing from 1 to 6. For more than 4 elements, it is no longer possible to build a completely nonredundant array. Instead, there are two alternatives: (1) the *restricted* array includes all multiples of a unit spacing up to some maximum, and then allows a redundancy for some of those multiples. For 5 elements, the two possible restricted arrays have element spacings of $*1*3*3*2*$ and $*1*1*4*3*$, both of which include all spacings from 1 to 9. (2) In the *general* array, all multiples of a unit spacing are included up to some intermediate spacing, then there is a gap where some of the multiples are not included. For 5 elements, these unrestricted arrays are $*3*1*5*2*$, which includes all spacings to 9, then there is a jump to 11, and $*4*1*2*6*$, which includes all spacings to 9, then there is a jump to 13. Results for larger arrays are tabulated in Leech and in Moffet. Leech also showed that for arrays with more than 4 elements, restricted and general configurations may be constructed with less than $\sim 30\%$ redundancy in the baselines.

In general, since gaps in the (u,v) plane cause sidelobe and sampling problems, the restricted minimum redundancy arrays are the better choice. Note that the redundancy cannot be broken while keeping uniform sampling – the emphasis really is that this is a *minimum* redundancy array. If one tries to push antennas around and use, say, $*1*2.7*3.3*2*$ instead of $*1*3*3*2*$, then some of the sampling is squeezed too tight and also some gaps are introduced. The minimum redundancy arrays are the arrays that cover the line uniformly, with the maximum number of spacings for a given number of elements and no gaps.

For two dimensional resolution, one could of course use a minimum redundancy linear array with earth rotation synthesis. As an alternative, one could apply the minimum redundancy technique to two or more linear arms in a two-

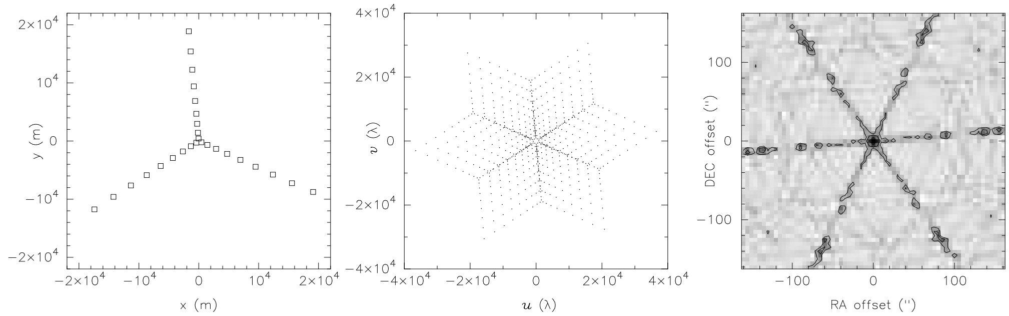


Figure 27–3. VLA Fourier plane coverage and beam in snapshot mode. The regularity found in the VLA snapshot coverage results in a different sort of grating response.

dimensional configuration. The current BIMA 10-element configurations were planned using restricted minimum redundancy arrays as the guides to station placement along East-West and North-South arms (Helfer & Welch 1997). Because minimum redundant linear arrays are used on both arms, the tracks for the separate arms have the maximum coverage, uniformly spaced, out to the longest baseline in units of one antenna diameter. The diagonals are also nonredundant and fairly uniformly separated. The net result is that the individual (u, v) tracks close on one another with a scan of about 8 hours, unlike the classic East-West arrays, which require a full 12 hour scan.

4.3. VLA-Y and GMRT-Y

The main benefit of the VLA's "Y" configuration is that it is a convenient 2-D arrangement of antennas which gives reasonable 2-D snapshot Fourier plane coverage. The BIMA and OVRO "T" arrays are similar in concept. The bad things about a "Y" or a "T" are that the regularity in the antenna directions along the arms will lead to a sort of grating response in the point spread function, and that it will take several hours of earth rotation synthesis for the Fourier samples to overcome this deficit.

The "Y" and "T" configurations are compromises: they seek to maintain the convenience of a 1-D array in reconfiguring the antennas, but would also like to get good Fourier plane coverage. As such, they are like arrays of fractal dimension 1.5: better than 1-D, not so good as fully 2-D arrays. Examples for the VLA's coverage and beams for a snapshot and a full track are shown in Figures 27–3 and 27–4.

The Giant Metrewave Radio Telescope (GMRT) in India has 14 inner antennas in a 1 km configuration and 16 in an outer configuration in the shape of an irregular "Y" (see Figure 27–5). The two configurations will often be used separately (the 30 m antennas were not designed for reconfiguration). The irregularity of the "Y" mainly stems from where land could be obtained, but it also produces a snapshot beam with 16 antennas which has much lower sidelobes than the VLA snapshot beam has with 27 antennas.

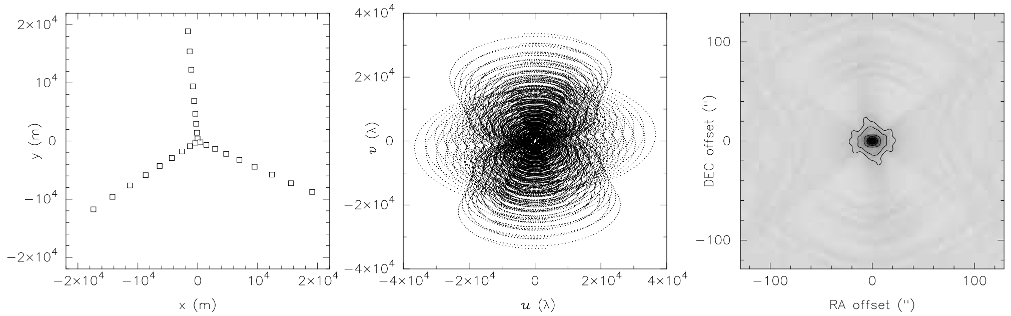


Figure 27–4. VLA Fourier plane coverage in a 12 hour integration. Long integrations remove the VLA's grating response. Note the skirt that surrounds this naturally weight PSF. This is due to the increased (u, v) density at short baselines found in the VLA's Fourier coverage.

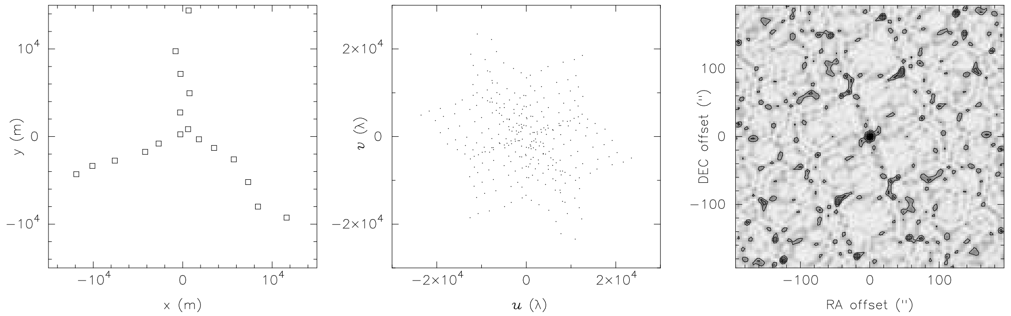


Figure 27–5. The GMRT outer configuration's snapshot coverage and beam. The irregularity in the antenna placement results in lower sidelobe levels.

4.4. Circular "Crystalline" Arrays

With the goal of achieving as uniform coverage as possible in the Fourier plane, Cornwell (1988) applied the optimization method of simulated annealing to maximizing the distance between (u, v) samples. In numerical optimization techniques, the trick is to avoid a solution which is stuck in some local minimum. In the statistical algorithm of simulated annealing, a user-defined "energy" function is minimized; to avoid local minima, the algorithm periodically boosts the energy of the system by a small amount, giving the system a chance to hop out of its local minimum and seek a new, lower energy solution.

In Cornwell's application, the energy function to be optimized is the logarithm of the distance between Fourier plane points. (The logarithm of the distance is used in order to emphasize close spacings.) Antenna configurations are generated at random, and the new configuration is always accepted if it minimizes the energy function (i.e., if it maximizes the logarithm of the distance between Fourier plane points). If the new configuration does not minimize the energy function, it is accepted anyway *if* it is not too much worse than the old configuration. It is this feature, where the algorithm is allowed to go uphill occasionally, that allows the solution to avoid a local minimum.

Cornwell constrained the antennas to lie within a circle, and the antennas "repelled" each other to lie on the circle. The snapshot Fourier plane distribu-

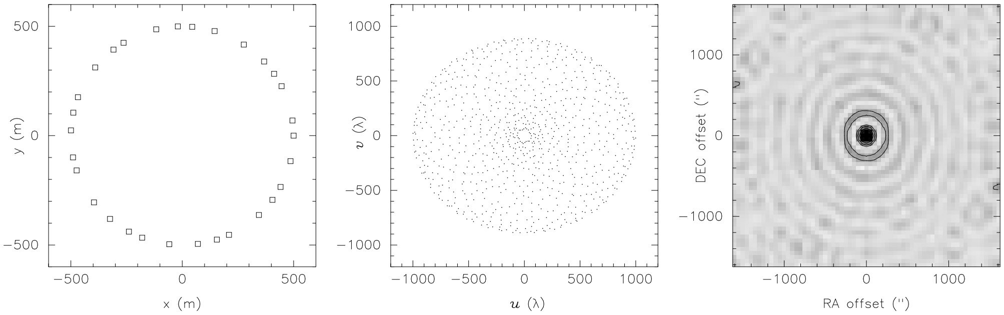


Figure 27–6. A crystalline array of 30 antennas and its snapshot coverage. The snapshot Fourier plane coverage at zenith is the autocorrelation of the array. This can be seen clearly in this (u, v) plot as the reader traces several little circles each of which are tangent to the origin and to the outer edge of the Fourier plane envelope. The sharp cutoff in the envelope results in nasty inner sidelobes in the beam.

tions are beautiful crystalline structures which exhibit bilateral symmetry (see Figure 27–6). While the antennas were allowed to be situated anywhere within the circle boundary, they migrated to the edge of the circle in the optimized array. This result suggests that uniform coverage in the (u, v) plane may best be achieved with ringlike arrays (see §4.5.).

While the simulated annealing approach to uniform Fourier coverage is successful for small numbers of antennas, the algorithm is computationally expensive; the time required per iteration goes like N^4. This is computationally prohibitive for arrays with large numbers of elements.

4.5. Uniform Fourier Plane Coverage: Elastic Net

In what is perhaps the most successful study of array design with the goal of achieving uniform Fourier plane coverage with a large number of antennas, Keto (1997) used a neural or elastic net algorithm in his application to the problem. Unlike the computationally expensive simulated annealing technique described above, which requires a compute time of N^4, the elastic net optimization is an N^2 operation.

The results of Keto's optimization for snapshot observations at the zenith show that the most uniform two-dimensional (u, v) coverage will be achieved if the antennas are located along the sides of a Reuleaux triangle (see Figure 27–7). Holdaway, Foster & Morita (1996) extended Keto's work to optimize over longer tracks at arbitrary declination and found that uniform coverage resulted best from some kind of closed figure, not necessarily a triangle, but rather ellipses or other "ringlike" arrays (see Figure 27–8), details of which depend upon the source declination and the length of the tracks. However, most of these configurations produced (u, v) distributions with very similar statistical properties.

Keto's Reuleaux triangle configurations, and ringlike configurations in general, yield fairly uniform (u, v) coverage plus a narrow peak at small spatial frequencies. They also offer the advantage of achieving the maximal sensitivity for the longest baselines, resulting in smaller naturally weighted resolution than other types of arrays with the same maximum baseline.

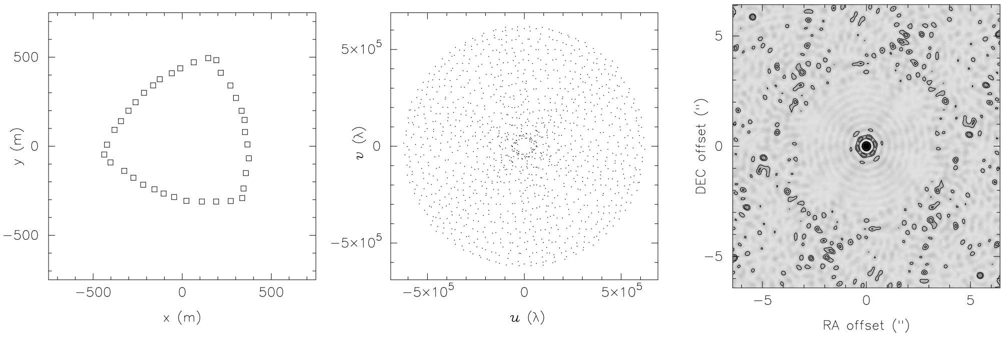


Figure 27–7. Configuration generated for a 36 element array from Keto's algorithm for a zenith snapshot. Note the large inner sidelobes.

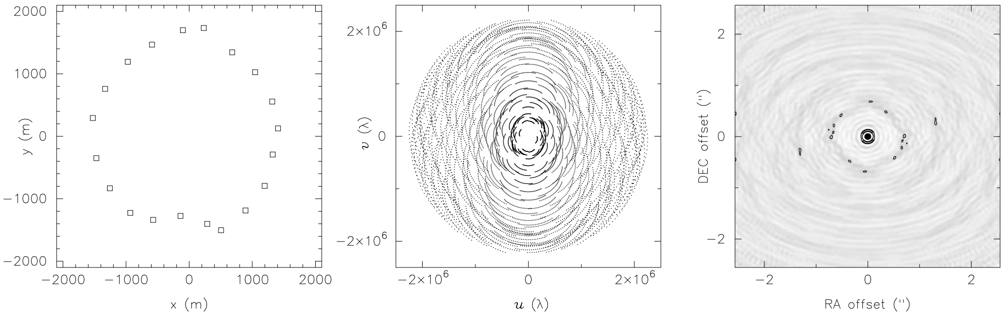


Figure 27–8. Configuration generated for a 20 element array from a modified Keto algorithm optimized for 4 hour tracks. The outer sidelobes are reduced by long tracks, but the inner sidelobes stay bad.

However, true uniform coverage in the Fourier plane has disadvantages as well: first, the sharp cutoff in (u, v) sampling at large spatial frequencies results in large (10-15%) sidelobes close to the central lobe of the synthesized beam (Holdaway 1996), which may complicate an image deconvolution and thereby lower its dynamic range (Holdaway 1996). Second, optimization techniques like the elastic net method used by Keto have so far tended to produce large diameters for the central hole in the Fourier plane coverage. It is probable that this problem can be alleviated to some extent, either by using nested rings or Reuleaux triangles, or by changing the optimization conditions to include some number of short baselines. However, the nested triangle approach destroys the uniform Fourier plane coverage. Third, unpublished simulations by Morita and by Holdaway show that it is the excess short spacing coverage which a ring array provides that is more responsible for high dynamic range in wide-field reconstructions than the uniform Fourier plane coverage. In other words, even if it were possible to get perfectly uniform Fourier plane coverage with a 36 element array, we probably would not want it.

4.6. Maximizing Brightness Sensitivity: Filled Array

Here, we move away from the goal of uniform Fourier plane coverage and instead turn our attention to the goal of maximizing the surface brightness sensitivity

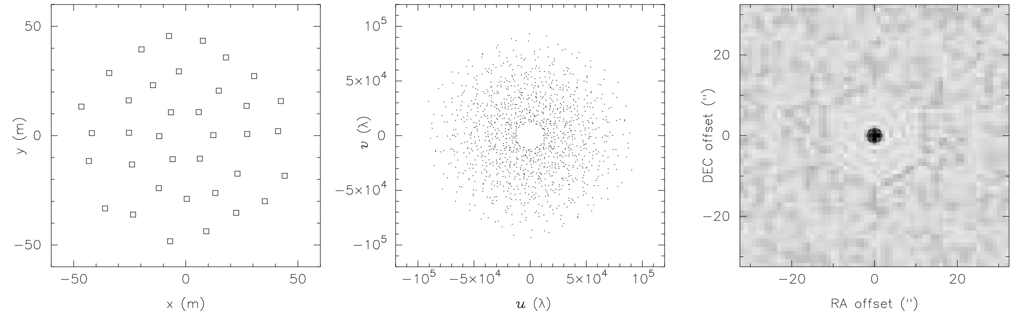


Figure 27–9. Snapshot coverage and beam of a 95 m filled configuration (Kogan 1998a). This is the closest thing we can get to the Fourier sampling of a large 95 m single dish. Except for the central hole, the Fourier plane density decreases nicely with (u, v) radius, making a nice beam.

of a configuration. This goal is the driving consideration for the construction of the compact D Array of the MMA. Surface brightness sensitivity is optimized by designing an array with the largest synthesized beam possible, which is achieved by having the shortest baselines possible. Optimizing the short baseline coverage is best achieved with a filled array, which produces a Fourier plane coverage that to first order is a linearly decreasing function of (u, v) distance. (Unpublished work and some MMA memos by R. M. Hjellming, by T. J. Cornwell, and by F. Owen may be credited as some of the early work which investigated filled arrays.)

The shortest baselines which may be achieved with a given array are limited strictly by the minimum safe distance which avoids mechanical collision of the antennas when pointing in arbitrary directions, which depends upon the antenna design, and less strictly by shadowing requirements (see §5.4.). Configurations with the highest density of the shortest baselines will be a hexagonally closely packed distribution of antennas, which results in an undesirably large grating response in the synthesized beam. Some degree of optimization is required for a compact array, trading off between good short baseline coverage and a large beam on the one hand and minimum synthesized beam sidelobes on the other. In the MMA application of 36 10 m antennas, with a minimum distance between antennas of 1.28 D, a sidelobe level of a few percent rms can be achieved with an array filling factor of 40% (Figure 27–9; Kogan 1998a). Such a filled compact array will result in complete instantaneous (u, v) coverage.

Kogan's algorithm is quite effective in generating filled arrays with minimum sidelobe levels (see Figure 27–9).

4.7. Minimizing PSF Sidelobes

An alternative philosophy for Fourier plane coverage has been studied by Kogan (1997, 1998a, 1998b, 1998c), who wrote an algorithm which produces antenna configurations which minimize the maximum sidelobe levels of the point spread function over a specified area in the image plane. Kogan's approach has the advantage of producing PSFs which should introduce fewer problems in image deconvolution. There is a pretty good, though not perfect, correlation between configurations which minimize the (u, v) gaps and which provide low sidelobes,

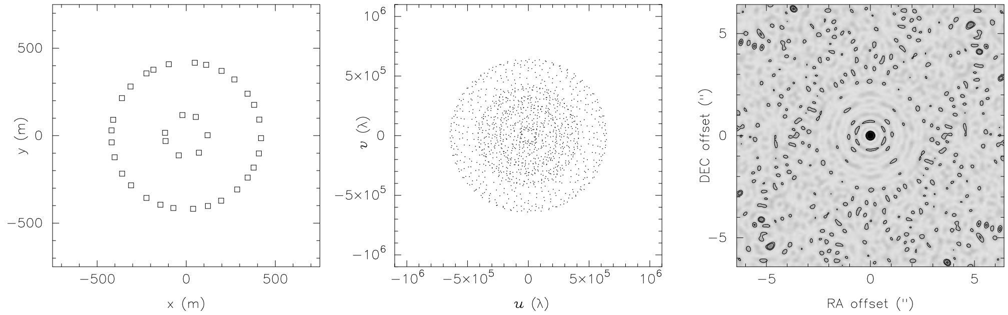


Figure 27–10. Snapshot coverage of a Kogan array optimized for minimum sidelobes. The region of the image plane which was specified for the sidelobe minimization is a circular region just inside of the little black sidelobes.

so that an array which minimizes sidelobes should have pretty uniform (u, v) coverage, but with a bit of natural tapering at long baselines to reduce the large inner sidelobes. Figure 27–10 shows the result of one such optimization in which the antennas were constrained to lie on two circles. In this application, the algorithm produces a fairly uniform, but gently tapered, Fourier plane distribution, where the coverage is the Fourier plane coverage of the outer ring (a large more or less uniform disk in the (u, v) plane), plus that of the inner ring (a small disk), plus that of the baselines made between the inner and outer rings (an annulus in the (u, v) plane). If the outer ring is 3 times larger than the inner ring (approximately true in this case), then the inner disk of (u, v) coverage and the middle annulus of (u, v) coverage will just touch. The inner disk and the middle annulus can be discerned in the (u, v) plot. In the beam, the circular region over which the sidelobes were optimized can be inferred from the surrounding high sidelobes.

An attractive feature of Kogan's approach is that it naturally shrinks the hole in the center of the (u, v) plane as the optimization extends over larger and larger regions in the image plane. This produces good coverage at short baselines in the (u, v) plane, which is one of the main shortcomings of Keto's uniform (u, v) coverage optimization.

Kogan's code can accept a variety of topographical constraints as inputs. Since the optimum in phase space is very broad, there will be many configurations which are approximately as good as each other. Even with apparently severe topographical constraints, Kogan's algorithm can produce configurations that are essentially as good as the unconstrained optimal configurations (see Figure 27–11). Furthermore, Kogan has investigated constraining the antennas to lie on concentric rings, and he has found that similarly low sidelobe levels can be obtained with such a constraint.

One disadvantage to Kogan's approach of minimizing the maximum sidelobe within some region of the point spread function is that rather large sidelobes can lurk just outside the region of optimization. It might be better to extend the region of optimization to the full width of the primary beam, and apply a weighting function which emphasizes the minimization of the close in sidelobes and gradually relaxes for the very far out sidelobes.

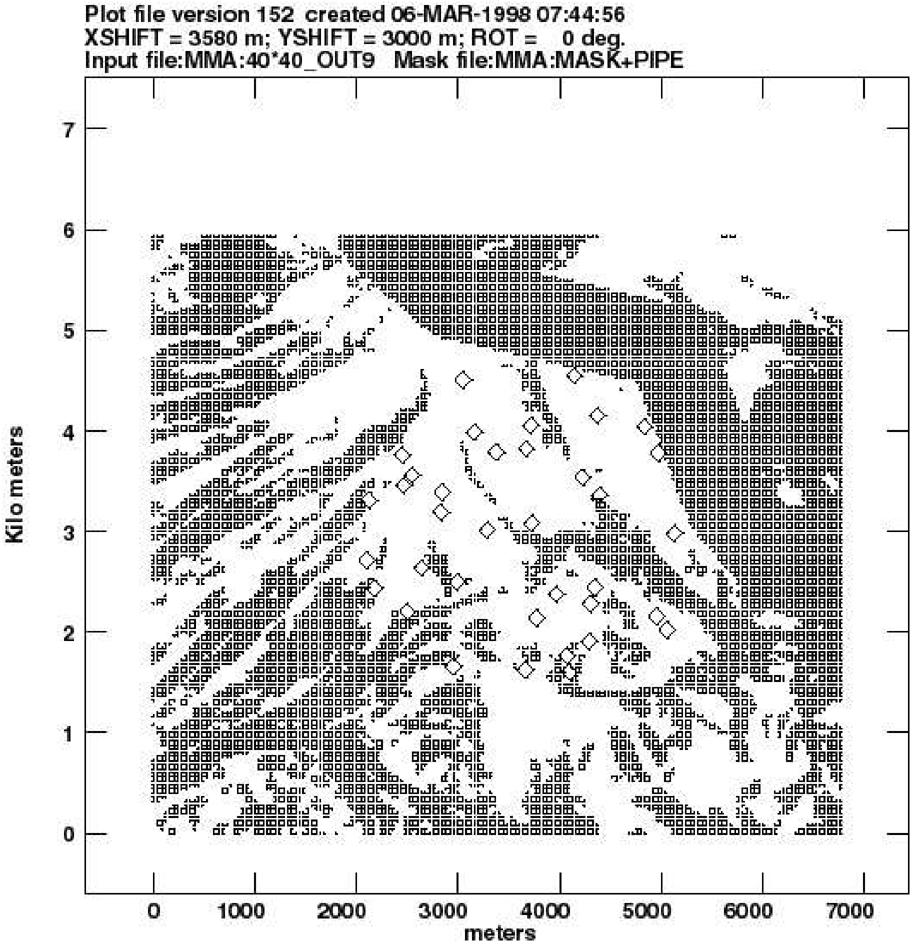


Figure 27–11. Example of a Kogan array which has been optimized subject to the constraint of avoiding any regions of surface slope more than 15% and a gas pipeline corridor on the MMA's future site at Chajnantor, Chile.

4.8. What Observations Do We Optimize Over?

The Keto, Cornwell, and Kogan arrays mentioned above were designed by algorithms. These computer algorithms seek to produce an array which optimizes some function of the (u, v) coverage or the point spread function. Each of these algorithms must perform its optimization for some specific observation. The easiest observation to optimize for is a snapshot at the zenith. The Cornwell (1988), Keto (1996), and Kogan (1998) algorithms in their native forms all perform their optimizations for a snapshot at the zenith. Kogan argues that an array which meets the minimum sidelobe criterion for a zenith at the snapshot is also optimal for a snapshot observation at transit for a source at any declination, but that an observation extended in hour angle will not necessarily have minimum sidelobes. Holdaway, Foster & Morita (1996) explored a recoded version of Keto's algorithm which optimized for extended hour angle track observations at an arbitrary declination, and found the resulting arrays to be different from

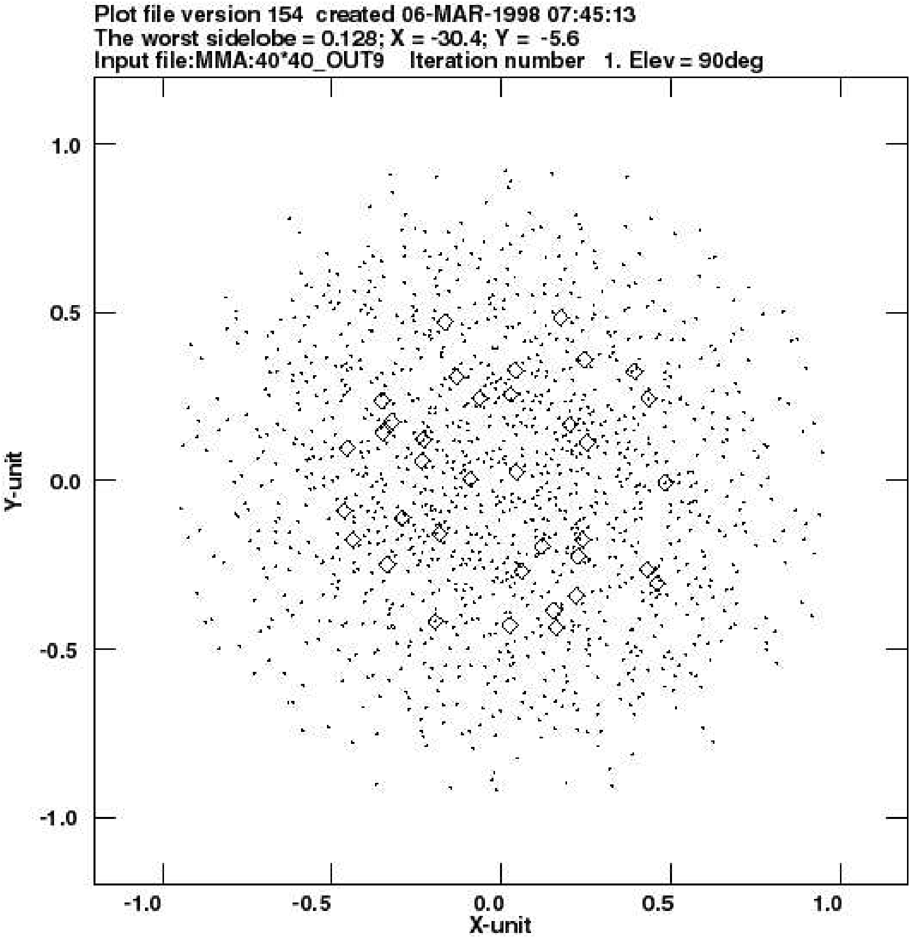


Figure 27–12. Corresponding (u, v) plane coverage for a snapshot with the topographically limited Kogan array.

those obtained from zenith snapshot optimizations. The gain in performing an optimization over long tracks and for a range of observations has not been quantified. It may improve the Fourier plane coverage negligibly over the coverage of a zenith snapshot optimized array.

For the MMA at least, the two compact arrays will provide nearly complete snapshot coverage, and the increasing system temperature due to high atmospheric airmass at low elevations will restrict the hour angle of observations to be within a few hours of transit, so optimizing for a transit snapshot is not inappropriate (Holdaway 1998b). The two more extended MMA configurations will require more time to fill in the (u, v) coverage, and optimization for long hour angle tracks over an appropriately weighted range of declinations may be important. Foster (1994) has demonstrated that the optimal N-S array elongation (i.e., the N-S elongation that results in the most circular beams, integrated over the sky) changes for snapshots and long tracks.

4.9. Simulations to Verify Image Quality

When we discussed work done with abstract Fourier plane coverages above, we found that the arrays which sought to produce those abstract Fourier plane coverages actually performed differently than the abstract Fourier plane coverages did. This is important: we should not base our array design blindly on an ideal, no matter how good that ideal sounds. The final test of an array is its imaging performance. Though time consuming, the quality of a proposed array configuration needs to be tested and compared with other competing array designs by using simulated data with representative source models and realistic errors, imaged with our current imaging algorithms. It is true that there may be some interaction between array configurations and imaging algorithms; we may be able to develop an algorithm which compensates for some currently undesirable feature of some (u, v) distribution. But we don't want to count our chickens with a hatchet.

5. Complicating Factors: Dealing with Physical Reality

5.1. Antenna Transportation

For multi-configuration arrays, the method of transporting the antennas may be a major influence on the array design. Antennas transported on railroad tracks favor a 1-D array. The VLA, the Owens Valley Radio Observatory (OVRO) millimeter array and the Nobayama Millimeter Array (NMA) have broken out of this mold in spite of their tracks; the "Y" and "T" configurations they use are still pretty easy to negotiate on rails. The Berkeley-Illinois-Maryland Association (BIMA) millimeter array has taken advantage of a free-standing antenna transporter to build stations that are offset from each other in both the North-South and East-West coordinate; the transporter can also be driven "off-road" to the distant (1 km) outrigger stations.

The MMA's set of compact configurations requires putting the antennas as close together as possible. This places strong maneuverability specifications on the antenna transporters.

5.2. Topography

The VLA is unencumbered by topography. It is desirable to be able to place any possible configuration on your site. This is not always possible. Even apparently severe topographical restrictions can turn out to be fairly minor if the optimization algorithm can optimize the array within these constraints. Kogan (1998) found that without any topographical constraints, the peak sidelobe level of a 36 element array was 0.126 in a region of the image plane out to 40 beam widths from the beam center. When topographical constraints due to arroyos on the MMA site were added, the peak sidelobe only increased to 0.128.

In addition, a computer readable digital elevation model (DEM) of the site, available from the USGS within the US, or from mapping firms in remote regions of other countries, such as the MMA site in northern Chile, can be used to determine the elevation limits placed on the antennas due to shadowing by the local terrain.

5.3. Overlapping Pads

Some capital costs and reconfiguration expenses can be cut by sharing pads between the different array configurations. For example, the VLA's "Y" lends itself to sharing stations quite nicely, and about half the antennas remain fixed in moves between adjacent arrays. Single ring arrays do not lend themselves to sharing pads between configurations. Double ring configurations could share considerably fewer than half the pads (the inner ring should have considerably fewer than half the antennas if we want to avoid highly centrally condensed Fourier plane coverage). Random arrays or donut arrays can share an intermediate number of pads between configurations. Conway (1998) proposed antenna layouts in self-similar spiral geometries, which allows for a large number of shared stations; this would allow for a continuously variable resolution, or for many fixed configurations with small resolution scale factors between them. A disadvantage of such an approach is that the arrays are essentially filled, which means that they do not have the uniform (u, v) coverage that is desirable especially for high-resolution configurations.

5.4. Shadowing

Geometrical shadowing of adjacent antennas must be considered for compact arrays. When antennas shadow each other, the illumination pattern on the dish changes, which will change the voltage pattern; the warm spillover increases, increasing the system temperature, and we are more susceptible to cross-talk (i.e., detecting signals originating in the electronics of the shadowing antenna). In our work, we assume that any antenna that is geometrically shadowed must be flagged as bad. There is an interaction between shadowing and a homogeneous mosaicing array's requirement for short spacings: we want the antennas close enough to give good short spacing coverage, but then the antennas will severely shadow each other when observing at lower elevation angles. This leads us to the requirement of multiple stretched compact configurations for different elevation ranges.

5.5. Sky Coverage and Stretched Configurations for Low Elevation

One important issue is sky coverage. This has bearings on where we locate our telescope on the earth. For example, a telescope situated near the equator will be able to see almost the full 4π steradians of sky (antenna elevation limits will clip a bit at the poles). Also, a telescope near the equator will see much of the sky at high elevation angles: half the sky can be viewed at elevation angles above 60 degrees, at which point there is little foreshortening of baselines and the synthesized beam is still nearly circular, and also the airmass is fairly low leading to near optimal opacities. A telescope located at a pole can see less than 2π steradians, and half the sky that it can see (i.e., one quarter of the sky) will be at elevation angles below 30 degrees, which will produce at least a 50% foreshortening of the baselines and an airmass of 2.0 or greater.

From the VLA, over 3π steradians are accessible. We would like to have good Fourier plane coverage over most of the accessible region, and we would also like to have a nearly circular beam as well. How circular the beam is depends upon the array configuration and site latitude, as well as the source declination and length of the tracks. Low elevation observations that produce highly elongated snapshot beams tend to get less elongated with longer tracks, but low declination sources are not up for very long either. The VLA's hybrid configurations, discussed in Alan Bridle's "guide", stretch the array in the N-S direction and provide a more nearly circular beam when observing southern sources.

Their are further complications for a mosaicing array. A single compact array designed for mosaicing which has good short spacing coverage will have a fairly limited sky coverage due to shadowing at low elevation angles. A stretched configuration which does not shadow at the lower elevation angles will have insufficient short spacing coverage at the higher elevations. A homogeneous mosaicing array will require three or four configurations to adequately cover the range of observable declinations (Holdaway & Foster 1996).

5.6. Atmospheric Influence on Sensitivity

It's a good thing to be able to observe your sources high in the sky for atmospheric reasons too: low elevation observations can suffer from high opacity and high system temperatures, which translate into low sensitivity, and phase errors also increase with decreasing elevation, though less severely than the opacity. This also favors an equatorial site, especially for high frequency observations where opacity and phase errors are more problematic. Requiring that a source be observed at reasonably high elevations will limit the potential hour angle tracks achievable by the array. For example, the MMA will most likely be used within a few hours of transit to maximize its sensitivity (Holdaway 1998b). This information must feed back into the array configuration design.

5.7. Paired Antenna Calibration

It is possible to build an array configuration with the antennas paired off with each other, using half the antennas to observe a calibrator source and the other half to observe a target source. The atmospheric phase fluctuations which the calibration subarray sees can be determined and then applied to calibrate the astronomical subarray. This scheme was once considered for phase calibration for the MMA, but it is not planned (see Lecture 28). This calibration method results in a loss of sensitivity (only half of the antennas are on source) and a loss of (u, v) coverage (one gets about 1/4 of the unique (u, v) samples, even if this calibration scheme is not required for a particular observation). In general, it is best to avoid flexibility limiting hardware solutions to problems that can be solved in other ways.

The proposed Japanese VLBI project VERA will measure Galactic dynamics via astrometrical observations. Since phase errors limit the astrometrical accuracy, the VERA array is proposed to have paired antennas to perform accurate phase calibration.

5.8. Redundant Arrays

In the early days of radio interferometry, calibration was problematic, and some arrays utilized redundancy to help calibrate the data. Redundancy is whenever the same physical baseline (length and orientation) is obtained from two different antenna pairs. Since the astronomical visibility is a function of baseline and is independent of the absolute location of the antennas, any differences in redundant baselines is due to the antenna gains. With enough redundancy, it is possible to fully calibrate the array to within a few offsets for sources which are bright enough to be detected with high SNR on each baseline within the atmospheric coherence time. Redundancy doesn't help the calibration of weak sources which are only detected after long integrations. The WSRT utilizes redundancy. The Nobeyama Radioheliograph is a modern array which uses redundancy for calibration; the dishes are so small that the only source in the sky bright enough to be seen is the sun, so it has no access to astronomical calibrators. Redundancy has its place, and supporters of redundancy are fairly enthusiastic. However, redundancy also takes away baselines that might otherwise be used to improve the instantaneous Fourier plane coverage. Furthermore, non-redundant synthesis arrays such as the VLA are generally able to calibrate well and to perform self-calibration when errors in the initial calibration limit the image quality.

6. Avoid Hardware Solutions to Software Problems

We need to design our telescopes within the framework of what we know we can do in terms of control software and imaging algorithms. For example, the pointing error and surface accuracy specifications of the MMA came out of the requirement that the mosaicing algorithm make high quality images. At this time, we cannot see any software solution to remove the effects of pointing errors or primary beam errors from mosaic images, so we have no choice but to design the MMA with these specifications in mind. In the lack of a software solution, we must design a hardware solution.

Here are some examples of hardware solutions to problems which can now be effectively solved in software:

- The NRAO 140 ft telescope in Green Bank was designed with an equatorial mount because it was feared that computers could not accurately control an AZ-EL mount radio telescope.

- Very large single dishes continue to be built to achieve "high sensitivity." Comparable sensitivity can be achieved with an array of smaller dishes. The smaller dishes operating interferometrically will have much lower systematic errors, and the total power observations are pushed to shorter baselines. The smaller dishes will also have a much wider primary beam, getting many resolution elements into each pointing on the sky. In order for a large single dish to keep up with an array, it needs to have large (~ 30 elements) multi-beam feeds at all frequencies (Holdaway & Rupen 1995).

- One of the benefits of linear E-W arrays is that all baselines remain coplanar as the earth rotates, and a sky projection which is tangent to the celestial pole eliminates the non-coplanar baseline problem which 2-D arrays such as the VLA must contend with at low frequencies. However, there is now an effective software solution to the non-coplanar baseline problem (see Lecture 19). Building a linear array is a hardware solution to what is now a software problem, paid for with the loss of snapshot imaging capability.

- Some E-W arrays have been designed so they can achieve complete, Nyquist sampled Fourier plane coverage (except for the central hole) in a 12 hour observation. Such complete Fourier plane coverage greatly reduces the need for deconvolution and can produce very high fidelity images. A 2-D array like the VLA will not obtain complete (u, v) coverage, but the effects of the holes in the (u, v) plane can, to a large extent, be removed by a software solution, deconvolution.

- One can imagine designing an array with very many, very small antennas to permit very large target sources to fit into the primary beam. As shown in the preceding section, this may not be optimal from a cost perspective. One can effectively synthesize a smaller dish by mosaicing, a software solution. One example of an array with very many, very small dishes is the Nobeyama Radioheliograph, which has 84 80-cm dishes, designed so the sun fits in its beam at the highest observing frequency. However, as this telescope monitors short time scale fluctuations of the sun's radio emission, it is not unreasonable to design this telescope with small dishes.

In each case, the hardware solution restricts the flexibility and power of the instrument. Hardware solutions to problems which can be effectively solved in software should generally be avoided.

7. Cases Studies

7.1. VLBA

VLBI arrays have traditionally been ad hoc arrays joining existing facilities which were not sited with VLBI observations in mind. The VLBA is the first VLBI array to be built with a configuration plan. It is a single configuration array intended to carry out imaging of the brightest continuum and spectral line astrophysical phenomenon at the highest resolution possible from U.S. territory. The higher observing frequencies used by the VLBA (15, 22, 43, and 86 GHz) required high elevation, dry sites to minimize the effects of atmospheric opacity and phase fluctuations. Another key consideration for the VLBA antenna sites was proximity to existing astronomical infrastructure, or at least proximity to airports and civilization. A common shortcoming in VLBI is the lack of short spacings, which prevents imaging of large scale jet structure. The VLBA addressed this by seeking to minimize the holes in a polar, logarithmic grid (Walker 1982). Since 10 antennas spanning a continent produces very sparse coverage, long tracks were used in the optimization. The Fourier plane coverage at several declinations were tested for each prospective array. The logarithmic grid makes smaller grid cells at short baselines and large ones at long baselines, biasing the VLBA configuration to include many more short baselines than long ones. In this way, the VLBA (u, v) coverage is somewhat "self-similar" with respect to tapering, and the imaging capability of the array is more or less independent of scale size.

To get the longest baselines, Walker chose Hawaii and sites in the far eastern US. Southern sites were emphasized to improve coverage of southern sources. As in any array optimization, the surface of the optimizing function in n-dimensional

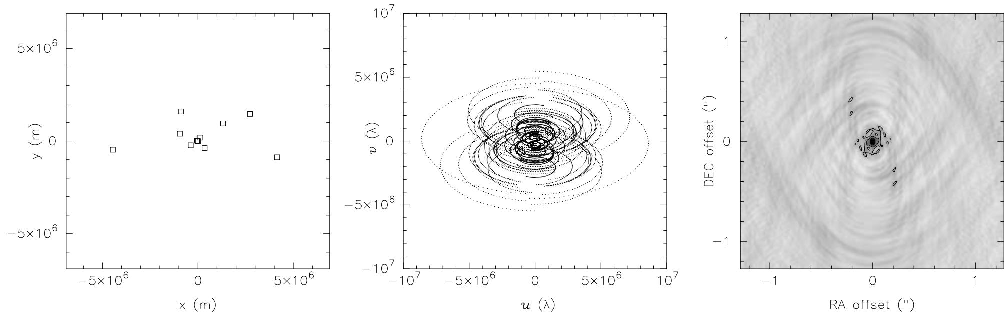


Figure 27–13. VLBA coverage and beam.

phase space has a very broad optimum: there are many, many configurations which are all approximately equally good. This allowed Walker a lot of freedom to move several of the VLBA stations to places like Owens Valley or Fort Davis where good astronomical infrastructure existed. Moving the other stations about corrected any deficit due to these choices. Finally, the optimization criteria using the polar logarithmic grid, along with the assumed use of the VLA with the VLBA, resulted in an array which was centrally condensed about the VLA. This centrally condensed array does result in a highly centrally condensed (u, v) coverage. In order to get the maximum resolution, uniform weighting (and a subsequent loss in sensitivity) is required.

To summarize: the VLBA had many constraints on it which did not translate simply into mathematical expressions. Logic and clear thinking drove the general form of the configuration design, and the centrally condensed (u, v) coverage criterion discriminated between good and poor array designs within the chosen general form. A sample VLBA (u, v) coverage and beam are shown in Figure 27–13.

7.2. MMA

We would have liked to show a case study of how the MMA configurations were chosen, but the ideas governing the MMA configurations are currently *diverging.* The MMA is currently forming an international partnership, and the joint array will consist of more, larger antennas than the originally proposed 40 x 8 m antennas. A larger, international pool of ideas and prejudices will go into rethinking the joint array design. This process has already begun (Kogan 1997, 1998a, 1998b, 1998c; Helfer & Holdaway 1998; Conway 1998; Webster 1998a, 1998b), but at this time we still don't know the dish diameter or the number of antennas! Among the issues that the international array configuration will need to address are:

- point source sensitivity
- the need for high quality mosaics and high surface brightness sensitivity.
- high resolution (maximum baseline of 10 km).
- simple reconfiguration operations at the remote 5000 m Chajnantor site.

- low sidelobes for good image reconstruction.
- flexibility in the choice of the desired Fourier distribution.
- local topographical features such as arroyos and shadowing from mountains impact the choice of antenna and array location.

Many of these issues have already been discussed in this lecture.

With so many antennas there is not a great amount of difference between the "best" and the "worst" possible arrays. And if we design a less than optimal array, there are still some tricks the astronomer may use to improve the situation.

8. Tools for Fixing Poorly Designed Arrays

8.1. Deconvolution

Array configurations which sample the Fourier plane rather poorly will have point spread functions with rather large sidelobes, resulting in low dynamic range images. This book has instructed us in how to fix this problem via deconvolution. However, deconvolution is not perfect. In general, the poorer the (u, v) sampling, or the larger the sidelobes in the point spread function, the larger the errors in the result of the deconvolution. The details of this statement depend strongly on the source structure being imaged.

8.2. Weighting Schemes

Configurations with centrally condensed Fourier plane coverage, like the VLA and the VLBA, result in beams which are much broader than λ/b_{max} (i.e., too much short spacing information results in too wide a beam). In addition, the VLA's naturally weighted beams have a wide pedestal that is sometimes problematic in image interpretation or deconvolution. The beam can be greatly improved by reweighting the Fourier data. Uniform and super-uniform weighting result in higher resolution beams, but at great expense in sensitivity for arrays with very centrally condensed coverage. Arrays with fairly uniform Fourier plane coverage see very little difference between natural and uniform weighting since they have very little redundancy, or the redundancy is spread out equally throughout the Fourier plane. Robust weighting (Briggs 1995) is a good compromise between good resolution, low sidelobes, and high sensitivity.

8.3. Changing the Way Existing Stations are Used

After it was commissioned, the Australia Telescope (AT) redesigned its configurations using the existing antenna stations. Some arrays, such as the Nobeyama Millimeter Array and the IRAM Plateau de Bure Interferometer, were built with a great many stations, and practical configurations have been worked out using a subset of the available stations. At the VLA, hybrid or N-S stretched arrays using existing station pads were not originally planned. The VLA is now using existing station pads from smaller configurations to help fill the unfortunately large hole in the center of the three larger arrays.

8.4. Building New Stations

If the requirements of actual observing are sufficiently different from the requirements envisioned when the configurations were designed, it may be necessary to build new stations. For example, the VLA could have a most compact configuration which has shorter shortest spacings and a factor 10 higher filling factor than the D array has, resulting in improved short baseline coverage for mosaicing and higher brightness sensitivity observations. This is one possible option for the VLA upgrade.

8.5. Multifrequency Synthesis

The VLA was not designed with snapshot observations in mind, so it should come as no surprise that the VLA's naturally weighted snapshot point spread function has very large sidelobes, some as large as 40%. The sidelobes can be decreased by improving the (u, v) coverage, either by earth rotation synthesis, or by observing at multiple frequencies within the band of interest. Observing at different frequencies changes the radial (u, v) distance, which is measured in wavelengths. A simple multifrequency approach has been utilized by the NRAO VLA Sky Survey (Condon et al. 1998), among other observers. This technique works only for continuum observations.

9. General Advice

A few bits of advice for all you home radio astronomers designing array configurations in your back yards:

- Design your array configuration for flexibility.
- Seek to solve problems in software or in observational strategy rather than hardware, which usually limits flexibility.
- Listen to the understood scientific requirements on the telescope, but not to the exclusion of flexibility to meet unknown scientific requirements which will only become apparent after the telescope has been completed.
- When in doubt, build lots of antennas. Everything works better with lots of antennas, except perhaps correlators and data reduction computers.

Acknowledgments. Many, many people have contributed to the understanding of how to design interferometric arrays, but we'd like to give special thanks to Frazer Owen, Tim Cornwell, and Scott Foster for their assistance in this work.

References

Braun, R. 1993, VLA Scientific Memo 165, NRAO.

Briggs, D. S. 1995, BAAS, 187, 112.02.

Condon, J. J., Cotton, W. D., Greisen, E. W., Yin, Q. F., Perley, R. A., Taylor, G. B., & Broderick, J. J. 1998, *AJ*, 115, 1693.

Conway, J. 1998, MMA Memo 216, NRAO.

Cornwell, T. J. 1988, *IEEE Trans. Ant. Prop.*, 36, 1165–1167.

Cornwell, T. J., Holdaway, M. A., & Uson, J. M. 1993, *A&A*, 271, 693–713.

Foster, S. M. 1994, MMA Memo 119, NRAO.

Helfer, T. T. & Welch, W. J. 1997, BIMA Memo 54.

Helfer, T. T. & Holdaway, M. A. 1998, MMA Memo 198, NRAO.

Hjellming, M. R. 1989, in *Synthesis Imaging in Radio Astronomy* eds. R. A. Perley, F. R. Schwab, & A. H. Bridle (San Francisco: PASP), 477–500.

Holdaway, M. A. 1994, VLA Scientific Memo 167, NRAO.

Holdaway, M. A. & Rupen, M. P. 1995, MMA Memo 128, NRAO.

Holdaway, M. A., Foster, S.M., & Morita, K.-I. 1996, MMA Memo 153, NRAO.

Holdaway, M. A. 1996, MMA Memo 156, NRAO.

Holdaway, M. A., & Foster, S. M. 1996, MMA Memo 155, NRAO.

Holdaway, M. A. 1997a, MMA Memo 177, NRAO.

Holdaway, M. A. 1997b, MMA Memo 178, NRAO.

Holdaway, M. A. 1998a, MMA Memo 199, NRAO.

Holdaway, M. A. 1998b, MMA Memo 201, NRAO.

Keto, Eric 1997, *ApJ*, 475, 843.

Kogan, L. 1998a, MMA Memo 217, NRAO.

Kogan, L. 1998b, MMA Memo 212, NRAO.

Kogan, L. 1998c, MMA Memo 202, NRAO.

Kogan, L. 1997, MMA Memo 171, NRAO.

Leech, J. 1956, *J. London Math. Soc.*, 31, 160.

Moffet, A. T. 1968, *IEEE Trans. Antennas Propagat.*, AP-16, 172.

Morita, K.-I. 1998, *in preparation.*

Rupen, M. P. 1997, VLA Scientific Memo 172, NRAO.

Rupen, M. P. 1998, VLA Scientific Memo 175, NRAO.

Walker, R. C. 1982, VLBA Memo 144, NRAO.

Webster, A. 1998a, MMA Memo 214, NRAO.

Webster, A. 1998b, MMA Memo 233, NRAO.

Wright, M. C. H. 1997, MMA Memo 180, NRAO.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

28. Millimeter Interferometry

C. L. Carilli
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

J. E. Carlstrom
University of Chicago, Chicago, IL 60637, U.S.A.

M. A. Holdaway
National Radio Astronomy Observatory, Tucson, AZ 85721, U.S.A.

Abstract.
Aspects of interferometry specific to observations at millimeter (mm) wavelengths are reviewed. The principal effects are: (i) the adverse effect of the troposphere on observed amplitudes and phases, and (ii) the demanding constraints on antennas and electronics. We begin with a short overview of some unique science that can be done with mm interferometry. We then discuss in detail the magnitude of the effect of the troposphere on system temperatures and absolute gain calibration, and its effect on interferometric phases, along with current methods to correct for these effects. We present constraints on antenna pointing and surface accuracy, and methods of determining these parameters. We then discuss some unique aspects of mm electronics. We conclude with a discussion the quantum limit to low noise receivers, and of the practical limits to the use of phase-conserving receivers at optical frequencies.

1. Science with mm Interferometers

Why build a mm interferometer? Interferometry at mm wavelengths offers many unique scientific capabilities.

Perhaps the most important of these capabilities is molecular line spectroscopy. Most of the low order rotational (electric dipole) transitions of cosmically abundant molecules fall in the mm spectrum. Table 28–1 lists a few of the more commonly observed transitions and their critical densities, where the critical density is defined as the density at which collisional excitation balances spontaneous de-excitation. The kinetic temperature required to excite these transitions lies in the range of a few to a few 10's of Kelvin[1], hence molecular spectroscopy with mm interferometers probes dense, cooler regions of the interstellar medium (ISM), such as dark molecular clouds and the early stages of star formation. Since dust obscures these regions at optical and near IR wavelengths, the detailed study of the early star formation process, and the study of the physics and chemistry of the dense ISM, has been the exclusive regime of mm astronomy (van Dishoeck et al. 1993). Hundreds of organic and inorganic molecules have been discovered in the ISM (Poynter & Pickett 1985), with some approaching the complexity of amino acids, the basic building blocks of life (Mehringer et al. 1997). An example of the plethora of lines available for study can be seen in the spectra of the Orion Molecular Cloud (Blake et al. 1987, Schilke et al. 1997), which show close to 1000 lines in the spectral ranges 208 GHz to 262 GHz and 330 GHz to 360 GHz (Figure 28–1), from molecules as simple as CO, to complex molecules such as CH_3CN. With the planned Mil-

[1] $\frac{h\nu}{k} = 14$ K at $\lambda = 1$ mm.

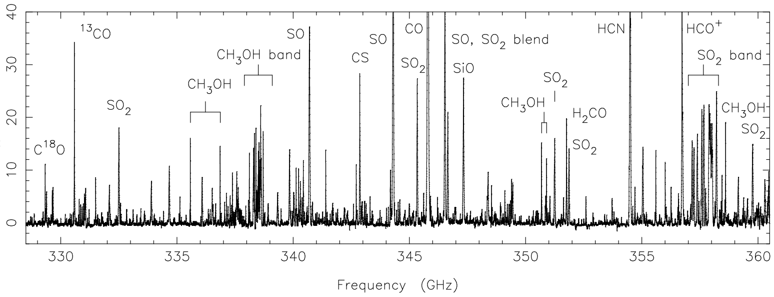


Figure 28–1. A spectrum of molecular emission from the Orion molecular cloud between 330 GHz and 360 GHz made with the Caltech Sub-mm Observatory (Schilke et al. 1997).

limeter Array (MMA) of the NRAO a spectrum of the Orion molecular cloud will be line-confusion limited in just a few minutes integration time.

A second area in which mm observations play a crucial role is in the study of thermal objects. In the Rayleigh-Jeans limit ($h\nu << kT$), the Planck emissivity function takes the form:

$$I_\nu = \frac{2kT\nu^2}{c^2} \text{ ergs cm}^{-2}\text{s}^{-1}\text{Hz}^{-1}\,. \tag{28–1}$$

Hence the thermal emissivity increases quadratically with frequency. A dramatic demonstration of this effect can be seen in Figure 28–2, which shows the predicted flux density for the starburst galaxy Arp 220 at various redshifts (Hughes 1998). The behavior of the emissivity curve is such that the increasing thermal dust emission with increasing frequency, i.e., the 'negative K-correction', offsets the distance losses, leading to flux densities that are roughly independent of distance, and even increase with distance over some redshift intervals. The sensitivity of the MMA, coupled with the expected space density of galaxies, are such that the MMA should detect continuum emission from about one 'normal' galaxy at z $\geq$ 1, such as the Milky Way, in every field observed at 230 GHz with an integration time of a few minutes or longer (Blain, Ivison, & Smail 1998, Brown 1996). Perhaps more importantly, mm observations are unaffected by dust obscuration. Current estimates of star formation rates in high redshift galaxies based on optical observations are highly suspect due to the possibility of dust obscuration (Madau et al. 1996). Millimeter and sub-millimeter observations provide the most direct, unobscured method for determining an accurate estimate for the star formation history of the universe (Hughes et al. 1998).

The high resolution provided by mm interferometers allows for a number of unique measurements. One such measurement is the study of thermal emission from optically thick regions, such as the surfaces of stars and proto-planetary disks (Levy & Lunine 1993, Wilner et al. 1996, Beckwidth & Sargent 1996). An observation of this type is shown in Figure 28–3, showing a spatially resolved

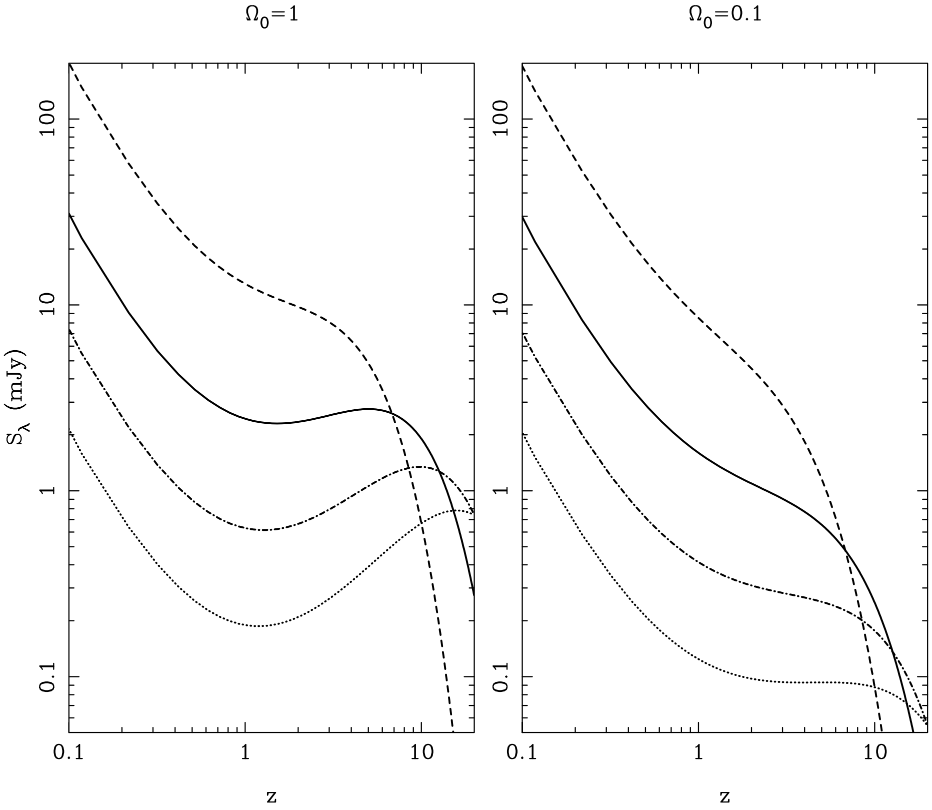


Figure 28–2. The expected flux density of the starburst galaxy Arp 220 as a function of redshift and observing frequency. Arp 220 has an $L_{FIR} = 3\times10^{12}$ $L_\odot$, and a massive star formation rate = 300 $M_\odot$ year^{-1}. The observed frequencies are: 450 μm (dashed line), 850 μm (solid line), 1350 μm (dashed-dotted line), and 2000 μm (dotted line). The left panel is for an Einstein-de Sitter universe. The right panel is for a low q_o universe (Hughes & Dunlop 1998).

image of the radio photosphere of the M Supergiant star Betelgeuse made with the VLA at 43 GHz. The resolution of the observation is 40 mas, and the measured diameter of the star is 80 mas (= 12 AU), with a brightness temperature of 3500 K (Lim et al. 1998).

Other important contributions from mm arrays will be imaging molecular emission from high redshift galaxies, and imaging molecular absorption by high column density quasar absorption line systems (Figure 28–4). Such observations provide detailed information on star formation and the dense ISM in nascent galaxies, at look-back times as high as 90% of the way back to the big bang. Thus far there have been six galaxies detected in CO emission at z > 2 (Barvainis 1998). In all cases the emission regions have characteristics similar to those seen in nearby nuclear starburst galaxies, with H_2 masses of order 10^{11} $M_\odot$. The emission occurs on scales of order a kpc or less, and the star formation rates are between 100 and 1000 $M_\odot$ year^{-1}.

To date, four molecular absorption line systems have been discovered at z $\geq$ 0.25 (Wiklind & Combes 1998). These systems are seen toward ‘red quasars’,

Table 28–1. Low Order Rotational Transitions of Simple Heavy Molecules

Molecule	J(1-0) GHz	J(2-1) GHz	J(3-2) GHz	n_{crit}[J(1-0)] cm^{-3}
CO	115.271	230.538	345.795	10^2 - 10^3
CS	48.991	97.981	146.969	10^3 - 10^4
HCN	88.631	177.260	265.886	10^5
HCO^+	89.188	178.375	267.557	10^5
SiO	43.122	86.243	130.268	10^3 - 10^4

i.e., quasars that are obscured in the optical due to dust associated with the absorbing gas. The molecular absorption lines provide detailed physical information on the dense, pre-star forming ISM in these galaxies, including isotope abundances, such as deuterium, and measurements of the excitation temperature of the molecular gas, which then constrains the evolution of the temperature of the cosmic microwave background radiation.

2. Problems Unique to mm Interferometry

The most important difference between mm and cm interferometry is the effect of the troposphere at mm wavelengths. The optical depth of the troposphere becomes significant below 1 cm, leading to increased system temperatures due to atmospheric emission, and increased demands on gain calibration due to variable opacity. Even more dramatic is the effect of the troposphere on interferometer phases. Variations in the tropospheric water vapor column density lead to variations in pathlength, and hence variations in interferometric phase. This can cause loss of visibility amplitude over the integration time ('coherence'), reduced spatial resolution ('seeing'), and pointing errors ('anomalous refraction').

Interferometry at mm wavelengths is demanding on both the antennas and the electronics. Typical mm-wave antennas are a factor of a few smaller in diameter than cm-wave antennas, but the frequencies are an order of magnitude or more higher, leading to stringent requirements on antenna pointing. A related requirement is the reduced field of view, thereby requiring 'mosaic' observations with many pointings for imaging celestial objects larger than an arcminute or so. The surface accuracy required increases linearly with frequency (Lecture 3), as does the required stability of the electronics (delays lines, oscillators, etc...).

MM interferometers require large bandwidths, up to a few GHz, for sensitivity, and for velocity coverage. The rest frame velocity covered, Δv, by a given observing bandwidth, $\Delta\nu_{\rm obs}$, behaves as:

$$\frac{\Delta v}{c} = \frac{\Delta\nu_{\rm obs}}{\nu_{\rm obs}}, \tag{28–2}$$

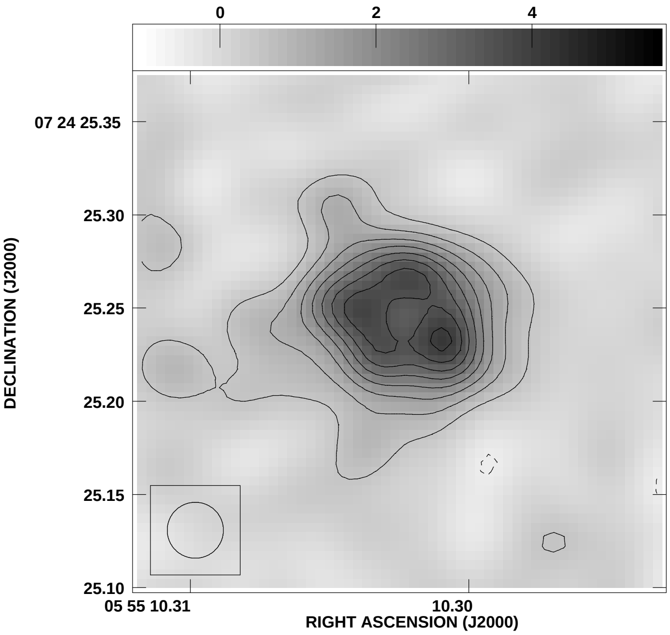


Figure 28–3. A VLA image of the radio photosphere of the M supergiant star Betelgeuse (α Orionis) at 43 GHz with a resolution of 30 mas. This image was made using the Fast Switching phase calibration technique with a total cycle time of 150 seconds. The measured diameter of the star is 80 mas, and the brightness temperature is 3500 K. The peak surface brightness is 4 mJy beam^{-1}, and the contour levels are $-1, -0.5$, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, and 4 mJy beam^{-1} (Lim et al. 1998).

where $\nu_{\rm obs}$ is the observing frequency. To obtain a spectrum covering a velocity range typical for a galaxy, $\Delta v \approx 300$ km s^{-1}, requires only 1.4 MHz bandwidth at 1.4 GHz, but 230 MHz bandwidth at 230 GHz.

Lastly, it is currently not possible to build low noise transistor amplifiers (LNAs) that work above 115 GHz (Maas 1992). Hence the receivers in mm interferometry are fundamentally different than in cm interferometry, requiring a mixer as the first stage. To achieve low noise characteristics then requires state-of-the-art designs for front-end electronics on mm telescopes.

3. The Troposphere

3.1. A General Description

The troposphere is the lowest layer of the atmosphere, extending from the ground to the stratosphere at an elevation of 7 to 10 km. The temperature decreases with altitude in this layer, clouds form, and convection can be significant. The troposphere is composed predominantly of N_2, O_2, trace gases such as water vapor, N_2O, and CO_2, and particulates such as liquid water and dust in clouds. Gen-

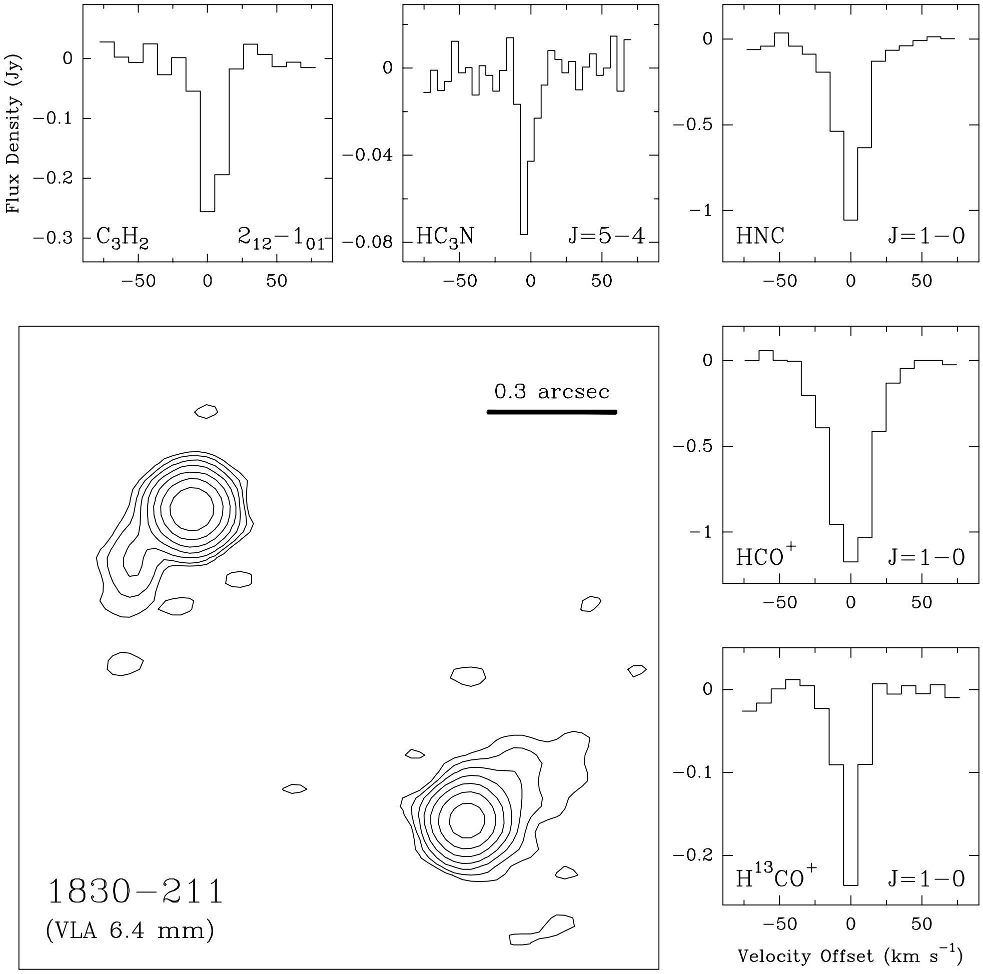


Figure 28–4. An image of the gravitationally lensed ('Einstein Ring') radio source PKS 1830-211 at 6.4 mm made with the VLA at 0.1″ resolution. The spectra show molecular absorption lines by gas in the lensing galaxy. Zero velocity corresponds to a heliocentric redshift of 0.88582, and the velocity scale is in km s^{-1}. Most of these transitions redshift into the VLA 43 GHz band. The absorption is seen toward the southwest continuum component which has a peak surface brightness of 1.2 Jy $beam^{-1}$ (Menten, Carilli, & Reid 1998).

erally, the troposphere becomes increasingly opaque with increasing frequency, mostly due to absorption by O_2 and H_2O.

Figure 28–5 shows models of the atmospheric transmission at cm and mm wavelengths for the VLA site at 2150 m altitude, and the planned millimeter array (MMA) site in Chile at 4600 m altitude. The plot shows a series of strong absorption lines including the water lines at 22 GHz and 183 GHz, and the O_2 lines at 60 GHz and 118 GHz, plus a systematic decrease in the transmission with increasing frequency between the lines. This 'pseudo-continuum' opacity is due to the sum of the pressure broadened line wings of a multitude of sub-mm and IR lines of water vapor. The plot for the MMA site at Chajnantor in Chile

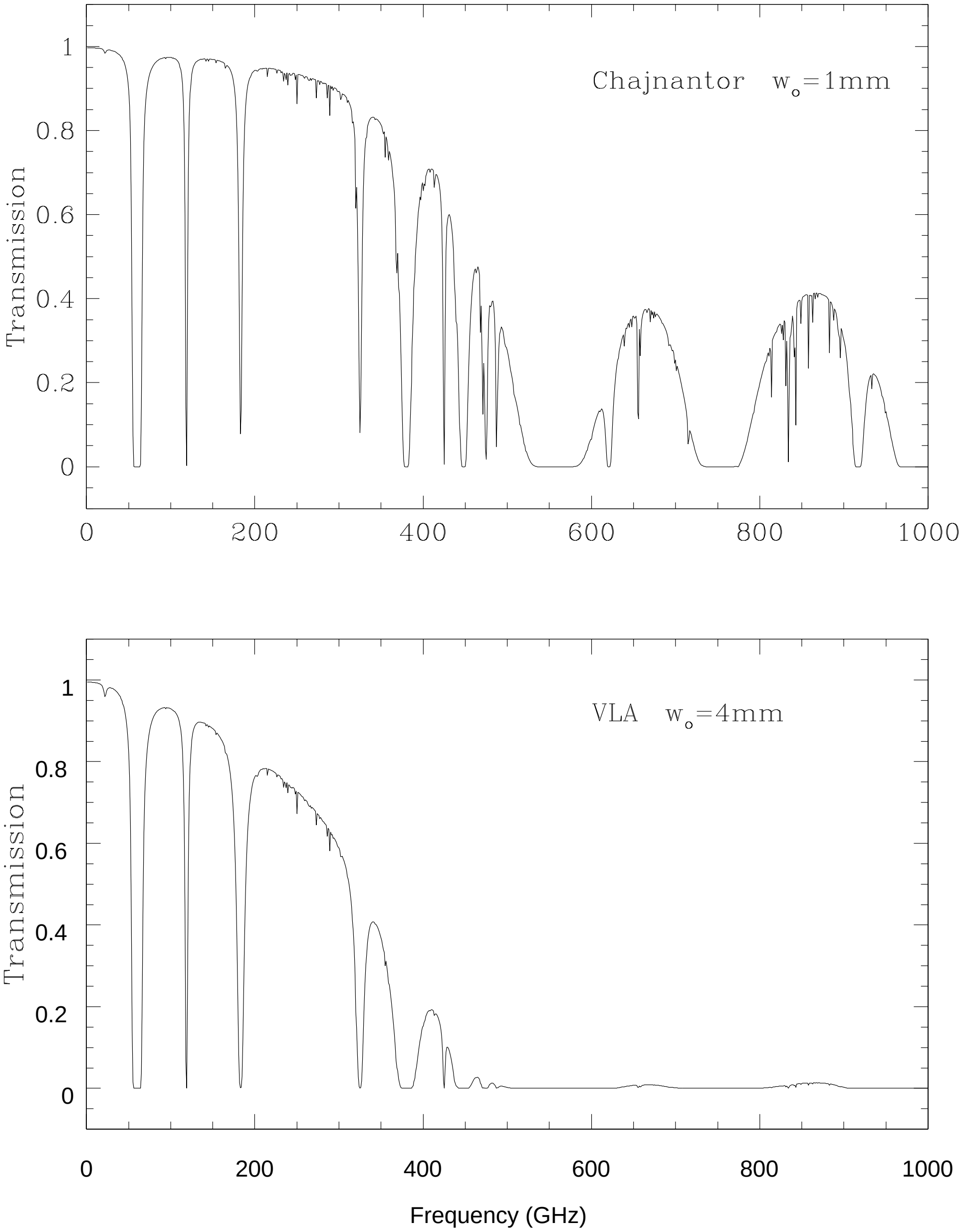


Figure 28–5. The upper frame shows the transmission of the atmosphere from 0 to 1000 GHz for the MMA site at Chajnantor in Chile assuming the typical value of w_o = 1 mm of precipitable water vapor, calculated using the Liebe atmospheric model (Liebe 1989, Holdaway & Pardo 1997). The lower frame shows the transmission of the atmosphere for the VLA site in Socorro, NM, assuming a value of w_o = 4 mm.

includes the typical value for the column density of precipitable water vapor, w_o = 1 mm, while the water vapor column for the VLA site is assumed to be w_o = 4 mm, where precipitable water vapor (PWV) = the depth of the water vapor if converted to the liquid phase.

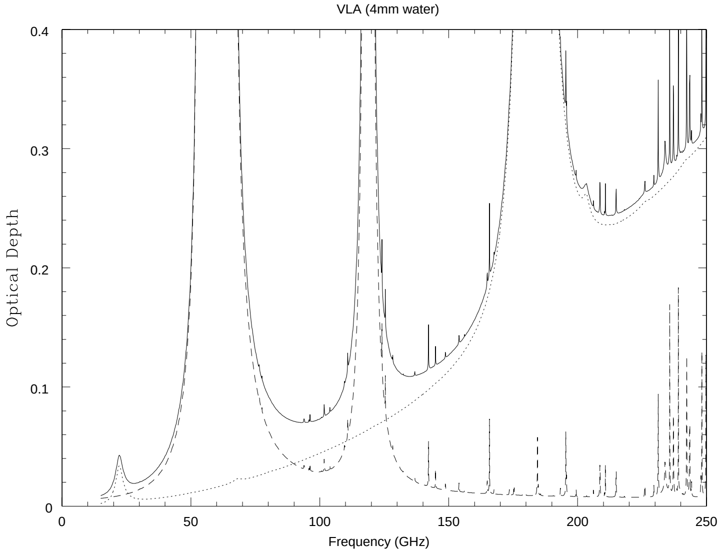


Figure 28–6. The optical depth for the VLA site assuming $w_o = 4$ mm. The solid line is the total optical depth. The dotted line is the optical depth due to water vapor. The dash line is the optical depth due to dry air (O_2 and other trace gases).

Figure 28–6 shows the relative contributions from water vapor and dry air (O_2 plus other trace gases) for the VLA site. Below 130 GHz both O_2 and H_2O contribute significantly. Above 130 GHz H_2O dominates the optical depth. An important aspect of the troposphere is that the total O_2 column is large, but relatively constant in time, while the water vapor column shows significant variations over time.

3.2. The Tropospheric Effect on $\mathbf{T}_{\rm sys}$

Let us consider a simple cascaded amplifier system. For a single amplifier the input astronomical signal, $S_{\rm in}$, and noise, N_1, is multiplied by the gain, G_1. The input noise includes contributions from the sky, plus ground pick-up by the telescope side-lobes, plus noise added by the electronics. The output from the amplifier is:

$$S^1_{\rm out} = G_1 \times (S_{\rm in} + N_1)$$

Adding a second amplifier with N_2 and G_2 then leads to:

$$S^2_{\rm out} = G_2 \times (S^1_{\rm out} + N_2) = G_2 \times [G_1 \times (S_{\rm in} + N_1) + N_2].$$

Each additional amplifier acts on the input signal and noise, such that the effect of the additional noise in the post-amplification electronics relative to the amplified signal (and input noise) is reduced by the gain factor of the amplifier. In other words, the 'effective input noise', $N^{\rm eff}_{\rm in}$ = the net noise contribution related back to the unamplified input signal, behaves as:

$$N^{\rm eff}_{\rm in} = N_1 + \frac{N_2}{G_1} + \frac{N_3}{G_1 G_2} + \ldots.. \qquad (28\text{–}3)$$

If the first amplifier has a high gain, then the system noise is set by the first amplifier, and whatever contributions occur before the amplifier.

The troposphere can be considered the first element in a cascaded amplifier system, with a 'negative gain' (i.e., loss), $G_{\rm atm}$, given by:

$$G_{\rm atm} = e^{-\tau_{\rm tot}},$$

where $\tau_{\rm tot}$ is the total atmospheric optical depth. The emission from the atmosphere acts as a black body in the telescope beam, with a noise temperature, $T_{\rm B}^{\rm atm}$, given by radiative transfer:

$$T_{\rm B}^{\rm atm} \ = \ T_{\rm atm} \ \times \ (1 \ - \ e^{-\tau_{\rm tot}}), \tag{28–4}$$

where $T_{\rm atm}$ is the physical temperature of the atmosphere. This equation is known as the 'radiometry equation' (Dicke et al. 1946).

The 'effective' system noise temperature, $T_{\rm sys}^{\rm eff}$, scaled to the top of the atmosphere (i.e., relative to the unattenuated celestial signal) becomes:

$$T_{\rm sys}^{\rm eff} = e^{\tau_{\rm tot}} \times [(1 - e^{-\tau_{\rm tot}}) \times T_{\rm atm} + T_{\rm rec}]\,, \tag{28–5}$$

where $T_{\rm rec}$ is the system temperature due to all other factors below the atmosphere. The first (multiplicative) term on the right hand side of equation 28–5, $e^{\tau_{\rm tot}}$, is due to the opacity, or 'negative gain', of the atmosphere, while the second term, $(1 - e^{-\tau_{\rm tot}}) \times T_{\rm atm}$, is the emission from the atmosphere. A rigorous treatment of the radiometry equation involves integration of the radiative transfer equations through the atmosphere using the vertical profiles of temperature and density, and using models for the spectral line shapes, such as is done in the atmospheric models in Figures 28–5 and 28–6.

As an example, the opacity at an elevation of 30° at the MMA site in Chile at 230 GHz is typically about 0.12, while the expected $T_{\rm rec} \approx 100$ K. The value of $T_{\rm atm}$ is about 270 K, leading to $T_{\rm B}^{\rm atm} \approx 30$ K. The effective system temperature becomes: $T_{\rm sys}^{\rm eff} \approx 130$ K. Hence the troposphere can make a significant contribution to the system noise temperature at 230 GHz, even for an excellent site such as Chajnantor. The rms water vapor fluctuations are typically: $w_{\rm var} \approx$ 0.043 mm, at the MMA site. At 230 GHz this implies fluctuations is the system temperature: $T_{\rm rms} \approx A_\nu \times w_{\rm rms} \times T_{\rm atm} \approx 1K$, or about 1% of the total system temperature.

3.3. $T_{\rm sys}$ and τ_o

Deriving $T_{\rm sys}$. We want to measure the effective system temperature referred to a source at the top of the atmosphere. Assume that the output power response of the whole system is linear with the input signal: $P_{\rm out}$ = m$\times(T_{\rm inp} + T_{\rm sys})$, where $\rm T_{inp}$ is some source, or 'load', above the atmosphere. This equation has two unknowns, $T_{\rm sys}$, and the scale factor m. To determine these parameters then requires two measurements of 'calibrated loads', $T_{\rm inp}$, situated above the atmosphere. One load can be the cosmic microwave background radiation, which is a black body at $T_{3K} = 2.73$ K that fills the beam. This is measured by observing blank sky. For the second load the standard method is to place a black body at atmospheric temperature, $T_{\rm BB} = T_{\rm atm}$, in front of the feed. At first inspection it would appear that this second load does not include the atmospheric emission and attenuation. However, consider the situation of placing the same load at the top of the atmosphere (and making it large so that it fills the beam). The power output of the system then includes the attenuated contribution from the black body = $T_{\rm BB} \times e^{-\tau_{\rm tot}}$, and the contribution from the atmospheric emission = $T_{\rm atm} \times (1 - e^{-\tau_{\rm tot}})$. In the case where $T_{\rm BB} = T_{\rm atm}$ the atmospheric emission exactly cancels the attenuation of the load, and one is left with a measurement that is equivalent to placing the load at the top of the atmosphere (Kutner & Ulich 1981). The effective system temperature is then:

$$T_{\rm sys} = (\frac{T_{\rm BB} - T_{3K}}{P_{\rm BB} - P_{3K}}) \times P_{3K} - T_{3K}$$

. Note that this relationship only holds for $T_{\rm BB} = T_{\rm atm}$, and that a few percent error is made by assuming an isothermal atmosphere.

Deriving $\tau_{\rm tot}$. Assuming an accurate value of $T_{\rm sys}$ can be measured, the standard method for determining the zenith opacity is to use 'tipping scans' (Butler 1996). This involves monitoring the system temperature while the antenna slews over a large angle in elevation ($\geq 50°$). Inverting the radiometry equation 28–4, and assuming the variable part of the measured $T_{\rm sys}$ is due to the increasing contribution from $T_{\rm atm}$ with air mass, then yields the value of $\tau_{\rm tot}$ at the zenith.

An alternative method for deriving $\tau_{\rm tot}$ is to track a celestial calibrator over a large range in zenith angle (Kogan 1997). The difficulty with this method is trying to separate the effect of increasing optical depth with air mass from the variation of antenna gain with elevation. In theory the functional forms are different, and hence separable. In practice, the parameters may be coupled, thereby requiring separate measurements of opacity and the antenna elevation gain curve.

3.4. Absolute Gain Calibration

Absolute gain calibration in the cm regime typically involves point source celestial calibrators (quasars) with non-variable, absolutely calibrated flux densities (Lecture 5). Unfortunately, no such sources are currently known at mm wavelengths.

One common method for deriving absolute flux densities at mm wavelengths is to observe the planets and/or the moon. The planets are roughly black bodies of known size and temperature, and hence act as a 'load' in the beam above the atmosphere (Bagri & Lillie 1993, Butler 1998). For example, Venus at a distance of 0.5 AU has a total flux density of 500 Jy at 43 GHz, and a diameter of 30″. Problems with this method are that current models for the mm emission from the planets are accurate to at best 10%, and many of the planets will be highly resolved by modern mm interferometers, thereby precluding standard interferometric gain calibration (Lecture 5). In this case single antenna calibration techniques could be employed. However, some of the planets will be partially resolved by the primary beam of the individual antennas, thereby requiring an accurate model for the primary beam shape. An interesting alternative is to use main sequence and/or giant stars (Yun et al. 1998). The sun at a distance of 10 pc has a diameter of about 1 mas and a flux density at 230 GHz of 1.3 mJy.

There are more than 100 of such stars accessible to the MMA, and the sensitivity of the MMA allows for accurate absolute gain calibration to be performed on such sources in a few minutes integration time.

If accurate measurements of $\mathrm{t}_{\mathrm{tot}}$, $T_{\mathrm{B}}^{\mathrm{atm}}$, and T_{sys} are available, plus an accurate model for the antenna elevation gain curve, it is possible to make an *a priori* absolute gain calibration independent of celestial sources. In essence, each antenna becomes an absolutely calibrated radiometer with respect to the true flux density of astronomical sources scaled to the top of the atmosphere. Such a method is employed at the VLBA at cm wavelengths, with a typical accuracy of a few percent (Lecture 22). However, for the VLBA the atmospheric terms are small, since the observing wavelengths are typically longer than 10 mm. For mm interferometers the atmospheric terms can dominate, and it is unclear that the atmospheric parameters can be monitored to the level required to obtain an absolute *a priori* gain calibration accurate to 1% (Yun et al. 1998).

3.5. The Mean Tropospheric Effect on Interferometric Phase

The troposphere has a non-unit refractive index, n. The refractive index is defined via the phase change experienced by an electromagnetic wave, ϕ_{e}, propagating over a physical distance, D:

$$\phi_e = \frac{2\pi}{\lambda} \times n \times D,$$

or in terms of 'electrical pathlength', L_{e}:

$$L_{\mathrm{e}} = \lambda \times \frac{\phi_{\mathrm{e}}}{2\pi} = n \times D\,.$$

The refractive index of air is non-dispersive (i.e., roughly independent of frequency,) except near the strong resonant water and O_2 lines, and is typically given as a difference with respect to vacuum ($n_{\mathrm{vacuum}} \equiv 1$), in parts per million, N, as (Waters 1967):

$$N \equiv (n-1) \times 10^6\,.$$

The index of refraction of air is typically separated into the dry air component, N_{d}, and the water vapor component, N_{wv}. These terms behave as (Waters 1967, Bean & Dutton 1968):[2]

$$\begin{aligned} N_{\mathrm{d}} &= 2.2 \times 10^5 \times \rho_{\mathrm{tot}} \quad \text{and} \\ N_{\mathrm{wv}} &= 1.7 \times 10^9 \times \frac{\rho_{\mathrm{wv}}}{T_{\mathrm{atm}}}\,, \end{aligned}$$

where ρ is the mass density in gm cm^{-3}. A detailed derivation of these relationships can be found in Thompson, Moran, & Swenson (1986).

[2] The inverse dependence on temperature for N_{wv} is due to the increased effect of collisions on the (mis-)alignment of the permanent electric dipole moments of the water molecules with increasing temperature (Waters 1967, Bean & Dutton 1968).

For water vapor alone, it can be shown that $\rho_{\rm wv} = \frac{w}{D}$. Using the equations above then leads to the relationship between the electrical pathlength, $L_{\rm e}$, and the precipitable water vapor column, w:

$$L_{\rm e} = 1.7 \times 10^3 \frac{w}{T_{\rm atm}} \approx 6.3 \times w$$

or:

$$\phi_{\rm e} \approx \frac{12.6\pi}{\lambda} \times w \qquad (28\text{–}6)$$

for $T_{\rm atm} \approx 270$ K (see also equation 5-8). This relation between electrical pathlength and precipitable water vapor column has been verified experimentally for a number of atmospheric conditions (Hogg, Guiraud, & Decker 1981).

A well known effect of the mean troposphere is on the expected position of the source. Snells' refraction law dictates that: $\frac{sin(i)}{sin(r)} = \frac{n_r}{n_i}$, where i and r are the incident and exit angles with respect to the normal. Under the simplifying assumption of a plane parallel atmosphere, it is straight forward to show that the change in expected position of a source, δ_θ, due to atmospheric refraction behaves as: $\delta_\theta \approx \delta_n \times tan(i)$ radians, where $\delta_n = n_r - n_i \approx 0.00025$. Hence for a source at an elevation, $i = 45°$, the antenna pointing must be adjusted by $\delta_\theta \approx 1'$. A related correction must be made for the delay off-set between antennas introduced by the troposphere relative to the geometric (i.e., vacuum) delay. This term is small (of order 10 radians at 230 GHz) for connected element interferometers with baselines $\leq$ few km, where the antennas are all observing at nearly the same elevation. But it can be large (of order 10^3 radians at 230 GHz) for VLBI observations, where the antennas can be observing at very different elevations. See Thompson, Moran, & Swenson (1986) for a detailed discussion of the tropospheric pointing and delay corrections for radio interferometers. These corrections are calculated, and removed, by the on-line system using local measurements of the atmospheric parameters (dew point, temperature, pressure), a model for the profiles of the quantities though the atmosphere, and a model for how n depends upon these parameters (Clark 1973a,b, 1974, 1987).

3.6. Phase Variations due to the Troposphere

Variations in precipitable water vapor lead to variations in the effective electrical path length, corresponding to variations in the phase of an electromagnetic wave propagating through the troposphere (Tatarskii 1978). Such variations are seen as 'phase noise' by radio interferometers. Since the troposphere is non-dispersive, the phase contribution by a given amount of water vapor increases linearly with frequency (except in the vicinity of the strong water lines). Hence, tropospheric phase variations are most prominent for mm and sub-mm interferometers, and can be the limiting factor for the coherence time and spatial resolution of mm interferometers (Hinder & Ryle 1971, Lay 1997a,b, Wright 1996).

The standard model for tropospheric phase fluctuations involves variations in the water vapor column density in a turbulent layer in the troposphere with a mean height, $h_{\rm turb}$, and a vertical extent, W, which moves at some velocity, v_a. This model includes the 'Taylor hypothesis', or 'frozen screen approximation', which states that: 'if the turbulent intensity is low and the turbulence is approximately stationary and homogeneous, then the turbulent field is unchanged over

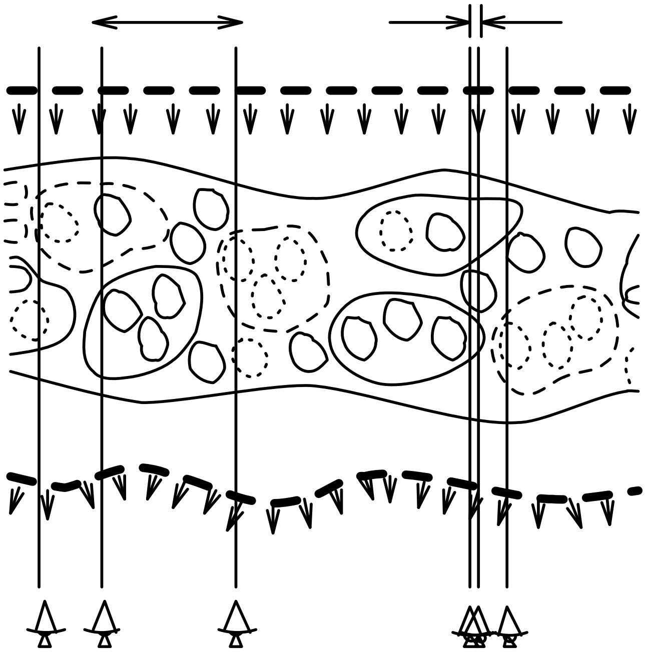


Figure 28–7. A schematic diagram of the effect of structure in the water vapor content of the atmosphere on different scales on interferometric phase. Circles of different sizes represent fluctuations in the water vapor content of the troposphere on various scales, including excesses (solid contours) and deficits (dotted contours). The phase of the incoming plane wave is distorted by the variations in the index of refraction due to variations in the water vapor content of the troposphere. The Taylor hypothesis implies that this phase screen advects across the array with the mean velocity of the winds aloft. Large scale fluctuations have the largest amplitude, but the effect on interferometric phase for closely spaced antennas is partially correlated. Smaller scale fluctuations have smaller amplitude, but are not correlated between antennas (figure from K. Desai 1998).

the atmospheric boundary layer time scales of interest and advected with the mean wind' (Taylor 1938, Garratt 1992). Under this assumption one can relate temporal and spatial phase fluctuations with a simple Eulerian transformation between baseline length, b, and v_a: $b = v_a \times$time. In the following sections we adopt a value of $v_a = 10$ m s^{-1}. This process is shown schematically in Figure 28–7.[3]

A demonstration of tropospheric phase fluctuations is shown in Figure 28–8 for observations with the VLA at 22 GHz. For these observations, two subarrays were employed, one observing the celestial calibrator 0423+418, and the second observing the calibrator 0432+416. The sub-arrays were 'inter-laced', meaning that every second antenna along each arm of the array observed a given source.

[3] Note that the theory remains valid even if a given fluctuation changes during the array crossing time, as long as the statistical properties remain unchanged, i.e., the statistics are stationary and homogeneous. In practice, observations at the VLA have shown that tropospheric phase fluctuations can often be 'tracked' across the array for many kilometers without dramatic changes.

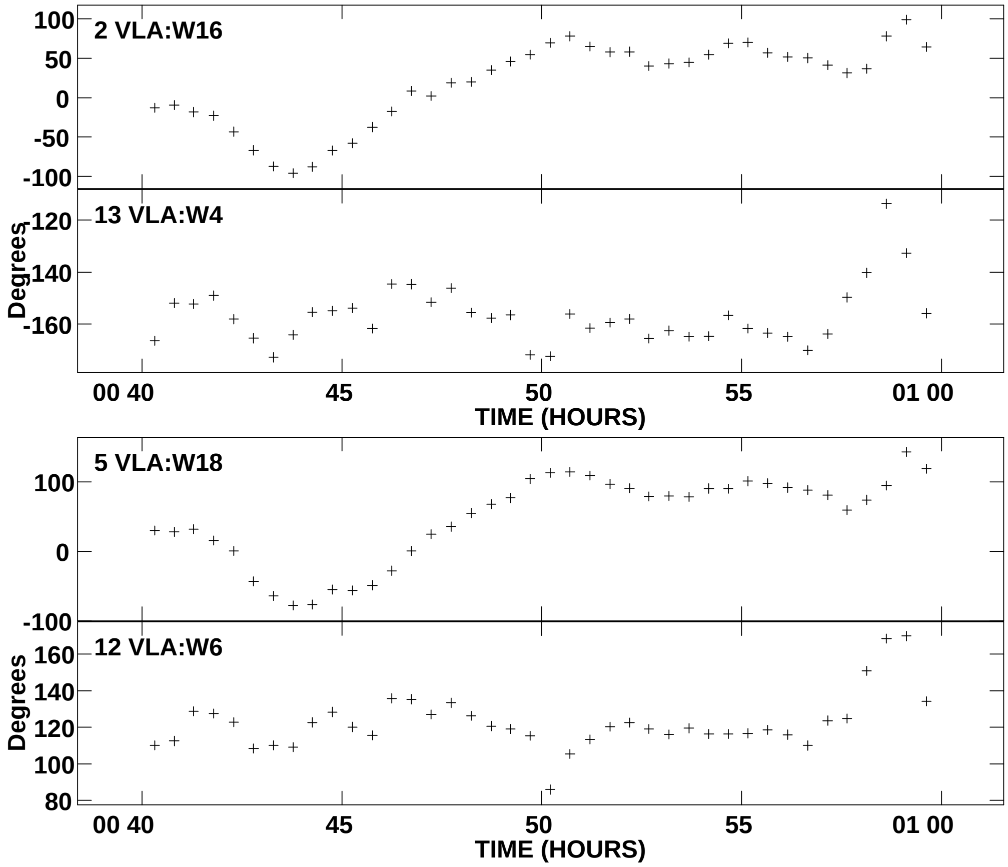


Figure 28–8. The top figure shows the antenna-based phase solutions vs. time for two antennas along the west arm of the VLA in a subarray observing the celestial calibrator 0423+418 at 22 GHz. The bottom figure shows the phase solutions over the same time for two antennas in a second subarray along the west arm observing the calibration source 0432+416. Antennas 5 (at station W18) and 2 (at W16) are at adjacent positions, and antennas 12 (at W6) and 13 (at W4) are adjacent. Note how the phases for the adjacent antennas track each other closely, even though the antennas in each adjacent pair are observing different sources. This is good evidence that the dominant contribution to the phase variations is the troposphere moving over the antennas.

The data in each subarray were self-calibrated with an averaging time of 30 sec. Figure 28–8 shows the antenna-based phase solutions from two pairs of neighboring antennas. The antennas at stations W16 and W4 were observing 0423+418 while the antennas at W18 and W6 were observing 0432+416. For adjacent pairs of antennas (W16-W18 and W6-W4) the temporal variations in the phase track each other closely. This close relationship for phase variations between neighboring antennas in the two different subarrays is the signature that the phase variations are primarily tropospheric in origin, and are correlated on relevant timescales and baseline lengths.

An important aspect of tropospheric phase fluctuations arising from the Taylor hypothesis is the relationship between the amplitude of the fluctuations and the time scale: large amplitude fluctuations occur over long periods and are partially correlated between antennas, while small amplitude fluctuations occur over short periods and are uncorrelated between antennas, depending on the baseline length. This effect can be seen in Figure 28–8 by the fact that the antennas at the outer stations (W14 and W16) show larger amplitude fluctuations relative to the inner stations (W4 and W6). This occurs because the phase solutions for each subarray are referenced to antennas at the center of the array, such that the reference antennas are within about 200 m of W4 and W6, but are separated from W16 and W14 by about 1000 m.

An example of what occurs in the image plane due to tropospheric phase fluctuations is shown in Figure 28–9. Observations were made of the celestial calibrator 2007+404 at 22 GHz with the VLA at a resolution of $0.1''$ (maximum baseline = 30 km) for a period of 1 hour. The data were self-calibrated using a long solution averaging time of 30 minutes, i.e., just a mean phase was removed from each half of the data. 'Snap-shot' images were then made from one minute of data at the beginning and end of the observation (upper left and upper right frames in Figure 28–9, respectively). Two important trends are apparent in these two frames. First, notice the positive-negative side-lobe pairs straddling the peak, indicative of antenna-based phase errors (Lecture 15). These image artifacts are due to phase fluctuations which arise in small scale water vapor structures in the troposphere that are not correlated between antennas. Second, notice that the peak in each image has shifted from the true source position. This position shift is due to phase fluctuations which arise in large scale water vapor structures in the troposphere that are correlated between antennas, i.e., a phase gradient across the array. These two frames are analogous to optical 'speckle' images, although the timescales for imaging are very different due to the larger scales involved in the radio.

The lower left frame shows the image of 2007+404 made using the full hour of data, but with a self-calibration averaging time of 30 minutes. The source appears extended in this image. The lower right frame shows the same image after self-calibration with an averaging time of 30 seconds, and in this case the source is unresolved. The lower left image has a peak surface brightness of 1.0 Jy beam^{-1}, an off-source rms noise level of 47 mJy beam^{-1}, and a total flux density of 1.5 Jy. The lower right image has a peak surface brightness of 1.6 Jy beam^{-1}, an off-source rms noise level of 5 mJy beam^{-1}, and a total flux density of 1.6 Jy. Not correcting for tropospheric phase noise in the lower left frame has: (i) increased the noise in the image, (ii) decreased the 'coherence' (i.e., lowered the peak surface brightness), and (iii) degraded the resolution ('seeing').

The important lesson from Figure 28–9 is that, while tropospheric phase errors can be quantified in terms of an antenna-based phase error, the errors are partially correlated between antennas on certain spatial and temporal scales, leading to positional shifts of sources as well as the standard positive-negative side-lobe pairs. Also, it is important to keep in mind that short baselines only 'sample' the power in the phase screen on scales of order the baseline length.

3.7. Root Phase Structure Function

Tropospheric phase fluctuations are usually characterized by the spatial phase structure function $D_\Phi(b)$,

$$D_\Phi(b) = \langle(\Phi(x+b) - \Phi(x))^2\rangle,$$

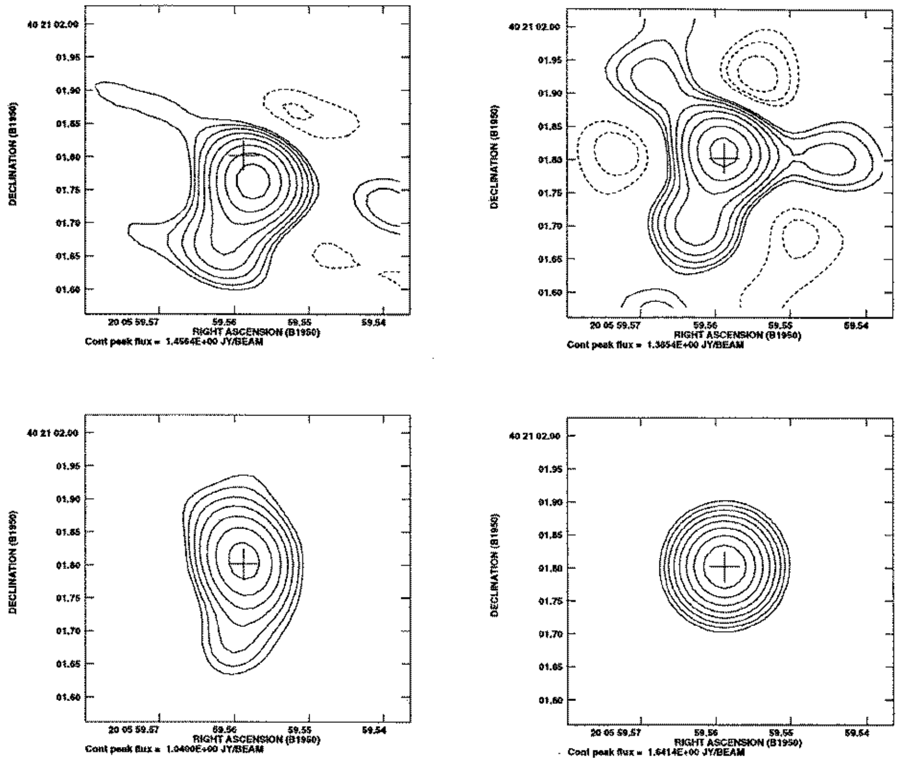


Figure 28–9. VLA images of the celestial calibrator 2007+404 at 22 GHz with a resolution of $0.1''$. The tick-marks on the declination axis are separated by 50 mas. The upper two frames show 'snap-shot' images made from one minute of data at the beginning (left) and end (right) of the one hour observation. These data were self-calibrated with an averaging time of 30 minutes, i.e., just a mean phase was removed for each half of the observation. The cross in each figure is a fiducial mark indicating the true position of the source. The lower left frame shows the image of 2007+404 made using the full hour of data with a self-calibration averaging time of 30 minutes. The lower right frame shows the same image after self-calibration with an averaging time of 30 seconds. The lower left frame has a peak surface brightness of 1.0 Jy beam^{-1}, and off-source rms noise of 47 mJy beam^{-1}, and a total flux density of 1.5 Jy. The lower right frame has a peak surface brightness of 1.6 Jy beam^{-1}, a noise of 5 mJy beam^{-1}, and a total flux density of 1.6 Jy. In all images the contour levels are a geometric progression in the square root of two, with the first level being 0.11 Jy beam^{-1}. Dotted contours are negative.

where b is the baseline length between two antennas, $\Phi(x + b) - \Phi(x)$ is the atmospheric phase difference measured between two antennas, and the brackets represent an ensemble average. Usually the ensemble average is replaced by a time average on one particular baseline. An interferometric array will sample the phase structure function at several baselines. For a single interferometer, the Taylor hypothesis (which asserts that temporal phase fluctuations are equivalent to spatial phase fluctuations) permits us to measure temporal phase fluctuations on a single baseline and translate these into the equivalent spatial phase structure

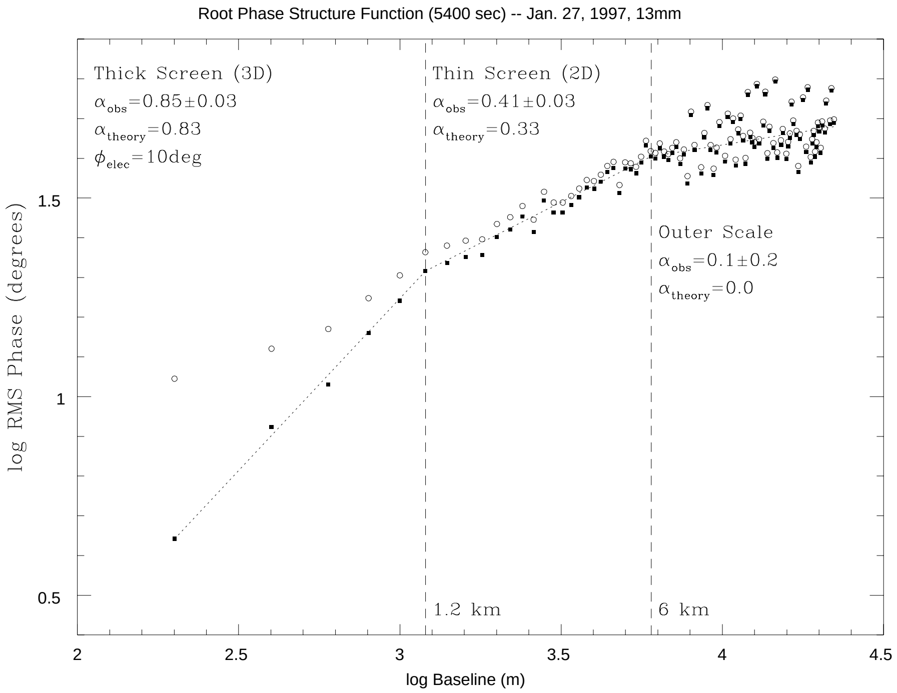


Figure 28–10. The root phase structure function from observations at 22 GHz in the BnA configuration of the VLA on January 27, 1997 (Carilli & Holdaway 1997). The open circles show the rms phase variations vs. baseline length as measured on the celestial calibrator 0748+240 over a period of 90 minutes. The filled squares show these same values with a constant noise term of 10° subtracted in quadrature. The three regimes of the root phase structure function as predicted by Kolmogorov turbulence theory are indicated.

function. In the following discussion we consider the square root of the phase structure function (the 'root phase structure function'), which corresponds to the rms phase variations as a function of baseline length (or time): $\Phi_{\rm rms} = \sqrt{D_\Phi}$.

Kolmogorov turbulence theory (Coulman 1990) predicts a function of the form:

$$\Phi_{\rm rms}(b) = \frac{K}{\lambda_{\rm mm}}\, b^{\alpha} \;\; \text{deg}, \tag{28–7}$$

where b is in km, and λ is in mm. The typical value of K is 100 for the MMA site in Chajnantor under good weather conditions, and $K = 300$ for the VLA site (Carilli, Holdaway, & Sowinski 1996, Sramek 1990).

Kolmogorov turbulence theory predicts $\alpha = \frac{1}{3}$ for baselines longer than the width of the turbulent layer, W, and $\alpha = \frac{5}{6}$ for baselines shorter than W (Coulman 1990). The change in power-law index at $b = W$ is due to the finite vertical extent of the turbulent layer. For baselines shorter than W the full 3-dimensionality of the turbulence is involved (thick-screen), while for longer baselines a 2-dimensional approximation applies (thin-screen). Turbulence theory also predicts an 'outer-scale', L_0, beyond which the rms phase should not increase with baseline length (i.e., $\alpha = 0$). This scale corresponds to the largest coherent structures, or maximum correlation length, for water vapor fluctuations in the troposphere, presumably set by external boundary conditions.

Recent observations with the VLA by Carilli & Holdaway (1997) support Kolmogorov theory for tropospheric phase fluctuations. Their result is reproduced in Figure 28–10, which shows the root phase structure function made using the BnA configuration of the VLA. This configuration has good baseline coverage ranging from 200m to 20 km, hence sampling all three hypothesized ranges in the structure function. Observations were made at 22 GHz during the night of January 27, 1997 using the VLA calibration source 0748+240. The total observing time was 90 min, corresponding to a tropospheric travel distance of 54 km, assuming $v_a = 10$ m s^{-1}. The open circles show the nominal tropospheric root phase structure function over the full 90 min time range.[4] The solid squares are the rms phases after subtracting (in quadrature) a constant electronic noise term of 10°, as derived from the data by requiring the best power-law on short baselines. The 10° noise term is consistent with previous measurements at the VLA indicating electronic phase noise increasing with frequency as 0.5° per GHz (Carilli & Holdaway 1996).

The three regimes of the structure function as predicted by Kolmogorov theory are verified in Figure 28–10. On short baselines ($b \leq 1.2$ km) the measured power-law index is 0.85±0.03 and the predicted value is 0.83. On intermediate baselines ($1.2 \leq\ b\ \leq 6$ km) the measured index is 0.41± 0.03 and the predicted value is 0.33. On long baselines ($b\ \geq 6$ km) the measured index is 0.1±0.2 and the predicted value is zero. The implication is that the vertical extent of the turbulent layer is: $W \approx 1$ km, and that the outer scale of the turbulence is: $L_0 \approx 6$ km. The increase in the scatter of the rms phases for baselines longer than 6 km may be due to an anisotropic outer scale (Carilli & Holdaway 1997).

3.8. Effects of Tropospheric Phase Noise

Coherence. Tropospheric phase noise leads to a number of adverse affects on interferometric observations at mm wavelengths. First is the loss of coherence of a measured visibility on a given baseline over a given averaging time due to phase variations. For a given visibility, $V\ =\ V_o \mathrm{e}^{i\phi}$, the effect on the measured amplitude of the due to phase noise in a given averaging in time is:

$$< V >= V_o \times < e^{i\phi} >= V_o \times e^{-\phi^2_{\rm rms}/2} \qquad (28\text{–}8)$$

assuming Gaussian random phase fluctuations with an rms variation of $\phi_{\rm rms}$ over the averaging time (Thompson, Moran, & Swenson 1986). For example, for $\phi_{\rm rms} = 1$ rad, the coherence is: $\frac{<V>}{V_o} = 0.60$, meaning the observed visibility amplitude is reduced by 40% from the true value.

Seeing. A second effect of tropospheric phase fluctuations is to limit the spatial resolution of an observation in a manner analogous to optical ‘seeing’, where optical seeing is due to thermal fluctuations rather than water vapor fluctuations.

[4] Note that the total observing time for calculating the rms phase fluctuations must be long for the larger configurations of the VLA, since the phase variations on a given baseline may have a significant, and perhaps even dominant, contribution from structures in the troposphere as large as five times the baseline length (Lay 1997).

Since interferometric phase corresponds to the measurement of the position of a point source (Lecture 2), it is clear that phase variations due to the troposphere will lead to positional variations of a source, and hence 'smear-out' a point source image over time (Figure 28–9). The magnitude of tropospheric seeing can be calculated by considering the coherence as a function of baseline length. Since the coherence decreases for longer baselines given an averaging time long compared to the array crossing time, the observed visibility amplitude decreases with increasing baseline length, as would occur if the source where resolved by the array. Using equation 28–7 for the root phase structure, and equation 28–8 for the coherence, the visibility amplitude as a function of baseline length becomes:

$$< V >= V_o \times \exp(-[\frac{K'b^\alpha}{\lambda\sqrt{2}}]^2) \tag{28–9}$$

Note that the exponent must be in radians, so $K' = K \times \frac{2\pi}{360}$. The baseline length corresponding to the half-power point of the visibility curve, $b_{1/2}$, then becomes:

$$b_{1/2} = (1.2 \times \frac{\lambda_{\rm mm}}{K'})^{1/\alpha} \ \ \mathrm{km}$$

For example, at 230 GHz using the typical value of $\alpha = 5/6$, and a typical value for K$'$ at the MMA site of 1.7, the value of $\mathrm{b}_{1/2} = 0.9$ km. This means that the resolution of the array is limited by tropospheric seeing to: $\theta_{seeing} \approx \frac{\lambda}{b_{1/2}} \approx 0.3''$ at 230 GHz. For average weather conditions, tropospheric seeing precludes diffraction limited resolution imaging for arrays larger than about 1 km at the MMA site in Chile, if no corrections are made for tropospheric phase noise. A rigorous treatment of tropospheric seeing, with predicted source sizes under various assumptions about the turbulence, can be found in Thompson, Moran, & Swenson (1986).

Two important points need to be remembered when considering tropospheric seeing. First is that the root phase structure function flattens dramatically on baselines longer than $\approx$ 1 km, such that the tropospheric seeing effectively degrades slowly with longer baselines. And second, although the coherence calculation above assumes an averaging time long compared to the array crossing time for the troposphere, the effect of seeing remains the same regardless of averaging time, since even for shorter integrations the position of the source is varying over time, thereby limiting the spatial resolution of the array when all the visibilities are used in imaging. This phenomenon can be seen in the lower left frame of Figure 28–9, in which the peak surface brightness is only 60% of the expected peak, but the total flux density averaged over the 'seeing disk' is 94% of the true value, i.e., the shortest baselines see the total flux density of the source even for long averaging times.

Anomalous Refraction. A final problem arising from tropospheric phase variations is 'anomalous refraction', or tropospheric induced pointing errors (Holdaway 1997, Butler 1997, Holdaway & Woody 1998). This effect corresponds to tropospheric 'seeing' on the scale of the antenna itself. Phase gradients across the antenna change the apparent position of the source on a time scale $\approx \frac{D}{v_a} \approx$ 1 second, for an antenna with a diameter D = 10 m. A straight-forward application of Snells' law shows that the effect in arcseconds should decrease with antenna diameter as $D^{-0.4}$. This is due to the fact that the value of α in the root phase structure function is less than unity, and hence the angle of the 'wedge' of water vapor across the antenna becomes shallower with increasing antenna size. However, in terms of fractional beam size, the effect becomes worse with antenna size as $D^{+0.6}$. For the 10m MMA antennas the expected magnitude of the effect at an elevation of 50° is $\approx 0.6''$.

3.9. 'Stopping the Troposphere': Techniques to Reduce the Effects of Tropospheric Phase Noise

An important point to keep in mind is that while tropospheric phase variations can be quantified in terms of a baseline-length dependent structure function, the errors are fundamentally antenna-based, and hence can be corrected by antenna-base calibration schemes, such as self-calibration or fast switching calibration.

Self-Calibration. A straight forward method of reducing phase errors due to the troposphere is self-calibration (Lecture 10). Self-calibration removes the baseline-dependent term in the root structure function, $\Phi_{\rm rms}(b)$, leaving the residual tropospheric phase noise dictated by the 'effective baseline': $b_{\rm eff} = \frac{v_a t_{\rm ave}}{2}$ = half the distance the troposphere moves during the self-calibration averaging time, $t_{\rm ave}$. The factor of two arises from the fact that the mean calibration applies to the middle of the integration time. The Taylor hypothesis dictates a relationship between temporal and spatial fluctuations such that the longer baselines will not sample the full power in the root phase structure function if the calibration cycle time is shorter than the baseline crossing time for the troposphere.

Of course, we would like to make $t_{\rm ave}$ as short as possible, but for a target source of some given brightness, we are limited in that we must detect the source in $t_{\rm ave}$ on each baseline with sufficient signal-to-noise ratio (SNR ≈ 2 for arrays with large numbers of antennas) to be able to solve for the phase. Hence, there will be sources which are so weak that they cannot be detected in a time short enough to track the atmospheric phase fluctuations. For the MMA at 230 GHz, self-calibration should be possible on fairly weak continuum sources (of order 10 mJy), with fairly short integration times ($\approx$ 30 sec), leading to residual rms phase errors $\leq 20°$. For the current VLA at 43 GHz the limit is 100 mJy sources with 30 second averaging times with residual rms phase variations of 10°. Self-calibration is not possible for weaker continuum sources, or for weak spectral line sources, or in the case where absolute positions are required. In these cases other methods must be employed to 'stop' tropospheric phase variations.

Fast Switching. Another method for reducing tropospheric phase variations is 'Fast Switching' (FS) phase calibration. This method is simply normal phase calibration using celestial calibration sources close to the target source, only with a calibration cycle time, $t_{\rm cyc}$, short enough to reduce tropospheric phase variations to an acceptable level (Holdaway 1992, Holdaway & Owen 1995, Carilli & Holdaway 1996, 1997). Holdaway (1992) shows that the expected residual phase fluctuations after FS calibration can be derived from the root phase structure function (equation 28–7), assuming an 'effective baseline length', $b_{\rm eff}$, given by:

$$b_{\rm eff} \;=\; d \;+\; \frac{v_a t_{\rm cyc}}{2} \qquad (28\text{–}10)$$

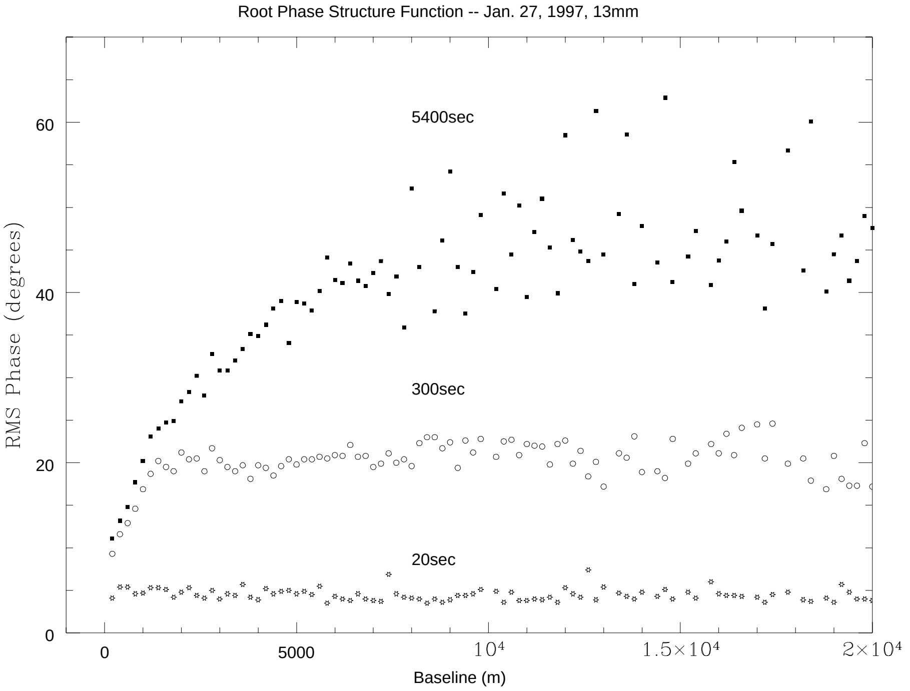


Figure 28–11. The solid squares are the same as for Figure 28–10, but now on a linear scale. These show the nominal root phase structure function from observations at 22 GHz with the VLA, as derived from a 90 min phase time series on the celestial calibrator 0748+240. The open circles show the residual rms phase variations vs. baseline length after self-calibrating the data with an averaging time of 300 seconds. The stars show the residual rms phase variations vs. baseline length after calibrating with a cycle time of 20 seconds.

where v_a = wind speed, and d = the physical distance in the troposphere between the calibrator and source. The FS technique will be effective for calibration cycle times shorter than the baseline crossing time of the troposphere $= \frac{b}{v_a}$. Moreover, a significant gain is made when $b_{\text{eff}} < 1$ km, thereby allowing for corrections to be made on the steep part of the root phase structure function, implying a timescale of 200 seconds or less for effective FS corrections. As with self-calibration, the calibrator source must be detected with sufficient SNR, and the cycle time must be short enough to track the atmospheric phase fluctuations.

A demonstration of the effectiveness of FS phase calibration is shown in Figure 28–11 for 22 GHz data from the VLA on baselines ranging from 100 m to 20 km (Carilli & Holdaway 1997). The solid squares show the nominal tropospheric root phase structure function averaged over 90 minutes (Figure 28–10). The open circles are the rms phases of the visibilities after applying antenna based phase solutions averaged over 300 seconds. The stars are the rms phases of the visibilities after applying antenna based phase solutions averaged over 20 seconds. The residual root structure function using a 300 second calibration cycle parallels the nominal tropospheric root structure function out to a baseline length of 1500m, beyond which the root structure function saturates at a constant rms phase value of 20°. The implied wind velocity is then: $v_a = \frac{2\times1500m}{300sec}$ $= 10$ m s^{-1}. Using a 20 second calibration cycle reduces b$_{\rm eff}$ to only 100 m, which is shorter than the shortest baseline of the array, and the saturation rms is 5°.

The important point is that, after applying standard phase calibration techniques on timescales short compared the array crossing time of the troposphere, the resulting rms phase fluctuations are *independent of baseline length for b > $b_{\rm eff}$*. The FS technique allows for diffraction limited imaging of faint sources on arbitrarily long baselines. Figure 28–3 shows an observation in which FS phase calibration was employed to produce a diffraction limited image of a faint astronomical source.

An important question to address when considering FS phase calibration is: are there enough calibrators in the sky in order to take advantage of a switching time as short as 40 seconds? This depends on the slew rate and settling time of the telescope, the set-up time of the electronics, the sensitivity of the array, and the sky surface density of celestial calibrators. Holdaway (1992) predicts that the mean distance between source and calibrator, θ, at 115 GHz will be:

$$\theta \approx 7 \times S_\nu^{0.75} \ \ \text{deg}$$

where S_ν is the source flux density in Jy. He shows that for calibrators with $S_\nu \geq 100$ mJy, and assuming a slew rate of 1 deg s^{-1}, the 40 element MMA should be able to employ FS phase calibration on most sources with total cycle times ≤ 20 sec, leading to residual rms phase fluctuations $\leq 20°$at 230 GHz, and on-source duty cycles $\approx 80\%$. One important practical problem is the lack of all-sky surveys at high frequency from which to generate calibrator source lists.

An important factor which will affect the ability of FS to correct for the atmospheric phase is the observing elevation. At low elevations, the radiation travels through more atmosphere and the phase fluctuations will be worse, the distance between the lines of sight will be large, and the opacity will be larger, making the sensitivity worse and the integration time on the calibrator will need to increase. However, at very high elevations, the time it takes for an AZ-EL telescope to slew in azimuth between the target and calibrator sources will increase due to the 1/cos(EL) effect. See Holdaway (1998) for a detailed study of these effects for the case of the MMA at the Chajnantor site.

Paired Array Calibration. A third method for reducing tropospheric phase noise for faint sources is paired antenna, or paired array, calibration. Paired Array (PA) calibration involves phase calibration of a 'target' array of antennas using a separate 'calibration' array, where the target array is observing continuously a weak source of scientific interest while the calibration array is observing a nearby calibrator source (Holdaway 1992, Counselman et al. 1974, Asaki et al. 1996, Drashkik & Finkelstein 1979). In its simplest form PA calibration implies applying the phase solutions from a calibration array antenna to the nearest target array antenna at each integration time. An improvement can be made by interpolating the solutions from a number of nearby calibration array antennas to a given target array antenna at each integration time. Ultimately, the discrete measurements in space and time of the phases from the calibration array could be incorporated into a physical model for the troposphere to solve for, and remove, the effects of the tropospheric phase screen on the target source as a function of time and space using some intelligent method of data interpolation, such as forward projection using a physical model for the troposphere and Kalman filtering of the spatial time series (Zheng 1985).

For simple pairs of antennas, the residual phase error can be derived from the root phase structure function with:

$$b_{\text{eff}} \approx d + \Delta b$$

where d is the same as in equation 28–10, and Δb is the baseline length between the calibration antenna and the target source antenna.

Figure 28–8 shows an observation for which PA calibration was implemented (section 3.6). Observations were made at 22 GHz using two 'inter-laced' subarrays observing two close calibrators, 0432+416 and 0423+416. Notice how the phase variations for adjacent antennas in the different subarrays track each other closely. This correlation between phase variations from neighboring antennas in different subarrays observing different sources implies that the tropospheric phase variations can be corrected using PA calibration.

In Figure 28–8 the temporal variations for neighboring antennas track each other well, but the mean phase over the observing time range is different between antennas. This phase off-set is due to the electronics and/or optics at each antenna, and should be slowly varying in time. Before interpolating phase solutions from the calibration array to the target array one must first determine, and remove, the electronic phase off-sets. This can be done by observing a celestial calibrator every 30 min or so. A demonstration of this process is shown in Figure 28–12.

A quantitative measure of the effects of PA calibration can be seen in the root phase structure function plotted in Figure 28–13. The open triangles show the tropospheric root structure function for the given observing day, as determined from the data with only the mean phase calibration (30 min averaging) applied. The shape of this function is well fit by a power-law in rms phase versus baseline length with index 0.65. The rms magnitude of 35°on a baseline of 1000 m is somewhat higher than the expected value of about 25°on a typical summer evening at the VLA (Carilli et al. 1996). The stars in Figure 28–12 correspond to the 'noise floor' for the phase measurements, as determined by calculating the root structure function from self-calibrated data. The solid squares in Figure 28–12 show the root structure function for the data with PA calibration applied. The residual rms values are about 10° on short baselines, and increase very slowly with baseline length. These data indicate a significant improvement in rms phase fluctuations after application of PA calibration for baselines longer than about 300 m.

The increased noise floor for the PA calibrated data relative to self-calibration indicates residual short-timescale phase differences which do not replicate between the target and calibration arrays. This noise floor is a combination of 'jitter' in the electronic phase contribution, and residual tropospheric phase noise as determined by b_{eff} above. Note that the residual noise floor increases slowly with baseline length. This is due to the logarithmically increasing separation between VLA antennas along the arm.

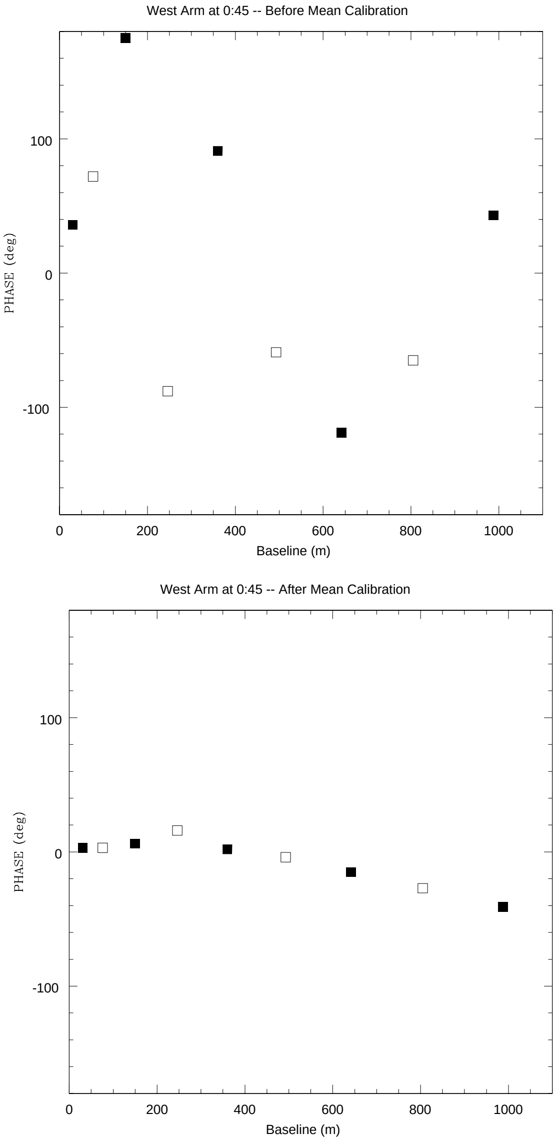


Figure 28–12. The top figure shows the antenna-based phase solutions at 22 GHz for antennas along the west arm of the VLA for a single 30 second observation. Two 'inter-laced' subarrays were employed. The solid squares are for antennas observing the celestial calibrator 0423+418 and the open squares are for antennas observing the celestial calibrator 0432+416. The bottom figure shows the same phase solutions after subtraction of the mean electronic phase averaged over 20 minutes. Note the random phase distribution along the west arm before the mean electronic phase is removed (upper frame), and the smooth phase gradient along the arm after this term is removed (lower frame).

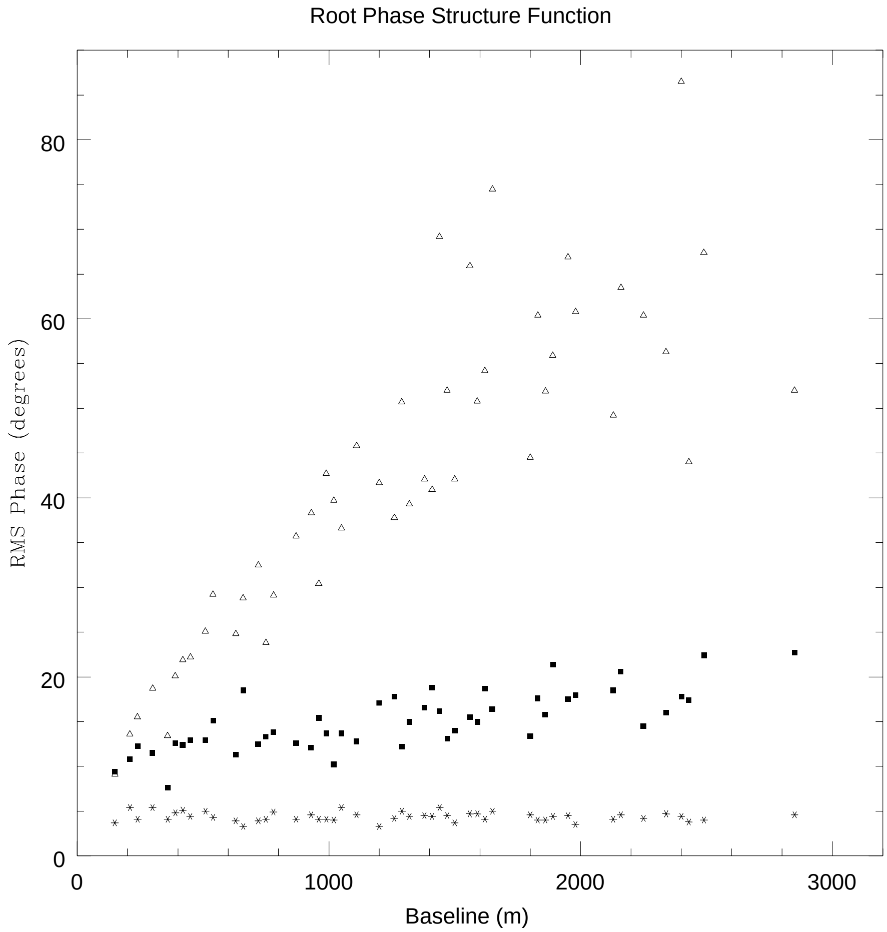


Figure 28–13. The open triangles show the root phase structure function for the data at 22 GHz on the 'target source' 0432+416 as determined from data with only a mean phase self-calibration applied (30 min average). The stars show the structure function after application of the self-calibration with 30 second averaging. The solid squares show the structure function after application of PA calibration with a 30 second averaging time, which entails applying the phase solutions from neighboring antennas observing the 'calibration source' 0423+418.

Fast Switching vs. Paired Array Calibration. Both FS and PA calibration have been demonstrated effective in reducing tropospheric phase noise for radio interferometers. We briefly discuss the relative advantages and disadvantages of the two techniques.

The advantages of FS are that: (i) FS uses the full array to observe the target source, and (ii) it removes the long and short term electronic phase noise along with the tropospheric phase noise. The disadvantages are that: (i) it places stringent constraints on telescope design in terms of slew rate, mechanical settling time, and electronic set-up time, and (ii) on-source observing time is lost due to frequent moves and calibration. For example, typical MMA observations using FS will have total cycle times of order 10 seconds, with fractional on-source times above 80% (Holdaway 1992).

There is a second factor which effects the FS observing efficiency: decorrelation caused by the residual phase errors (i.e., the small scale phase fluctuations which are too fast to calibrate via fast switching). We can try to make these residual phase errors smaller by reducing the cycle time, but this means we spend a smaller fraction of the time on source. Or we could increase the cycle time to optimize the fraction of the time spent on source, but at some point we lose sensitivity because our residual phase errors increase. Again, simulations for the MMA indicate that overall fast switching efficiencies (including both sensitivity losses due to decreased observing time and increased decorrelation) of about 0.75 or 0.80 should be possible for most observing frequencies and typical observing conditions at the Chajnantor site.

The advantages of PA calibration are that: (i) the 'target array' observes the source continuously, and (ii) the demands on the antenna mechanics and electronics are less stringent than for FS. The disadvantages are that: (i) the electronic phase noise is not removed, (ii) the geometry of the array must allow for neighboring antennas, even in large arrays, and (iii) the number of visibilities from the target source array decreases quadratically with the decreasing number of antennas. These latter two effects reduce significantly the Fourier spacing coverage for any given observation.

We envision that both FS and paired array calibration will be used at the MMA, depending on the configuration and the scientific requirements of a given observation.

Radiometry. By measuring fluctuations in $T_{\rm B}^{\rm atm}$ with a radiometer, one can derive the fluctuations in the column density of water vapor of the troposphere using the radiometry equation (Barrett & Chung 1962, Staguhn et al. 1998, Staelin 1966, Westwater & Guiraud 1980, Rosenkranz 1989, Welch 1994, Bagri 1994, Sutton & Hueckstadt 1997, Lay 1998). The relationship between electrical pathlength and water vapor column (equation 28–6) can then be used to derive the variable contribution from water vapor to the interferometric phase.

We assume that the atmospheric opacity can be divided into three parts:

$$\tau_{\rm tot} \ = \ A_\nu \times w_o \ + \ B_\nu \ + \ A_\nu \times w_{\rm rms}\,, \qquad (28–11)$$

where: (i) A_ν is the optical depth per mm of PWV as a function of frequency, (ii) w_o is the temporally stable (mean) value for PWV of the troposphere, (iii) B_ν is the total optical depth due to dry air as a function of frequency (also assumed to be temporally stable), and (iv) $w_{\rm rms}$ is the time variable component of the PWV of the troposphere. It is this time variable component which causes the tropospheric phase 'noise' for an interferometer. In effect, we assume a constant mean optical depth: $\tau_o \equiv A_\nu \times w_o + B_\nu$, with a fluctuating term due to changes in PWV: $\tau_{\rm rms} \equiv A_\nu \times w_{\rm rms}$, and that $\tau_o >> \tau_{\rm rms}$.

Inserting equation 28–11 into equation 4, and making the reasonable assumption that $A_\nu \times w_{\rm rms} << 1$, leads to:

$$T_{\rm B} \ = \ T_{\rm atm} \times [1 - e^{-\tau_o}] \ + \ T_{\rm atm} \times e^{-\tau_o} \times [A_\nu \times w_{\rm rms} \ + \ \frac{(A_\nu \times w_{\rm rms})^2}{2} \ + \ ...]. \qquad (28–12)$$

The first term on the right-hand side of equation 28–12 represents the mean, non-varying $T_{\rm B}$ of the troposphere. The second term represents the fluctuating component due to variations in PWV, which we define as:

$$T_{\rm B}^{\rm rms} \ \equiv \ T_{\rm atm} \times e^{-\tau_o} \times [A_\nu \times w_{\rm rms} \ + \ \frac{(A_\nu \times w_{\rm rms})^2}{2} \ + \ ...] \qquad (28\text{–}13)$$

At first inspection, it would appear that equation 28–13 applies to fluctuations in a turbulent layer at the top of the troposphere, since the fluctuating component is fully attenuated (i.e., multiplied by $e^{-\tau_o}$). However, for a turbulent layer at lower altitudes there is the additional term of attenuation of the atmosphere above the turbulent layer by the turbulence. It can be shown that the terms exactly cancel for an isobaric, isothermal atmosphere, in which case equation 28–13 is *independent* of the height of the turbulence.

Absolute radiometric phase correction entails measuring variations in brightness temperature with a radiometer, inverting equation 28–13 to derive the variation in PWV, and then using equation 28–6 to derive the variation in electronic phase along a given line of sight.

As benchmark numbers for the MMA we set the requirement that we need to measure changes in tropospheric induced phase above a given antenna to an accuracy of $\frac{\lambda}{20}$ at 230 GHz at the zenith, or $\phi_{\rm rms} = 18°$. This requirement inserted into equation 28–13 then yields a required accuracy of: $w_{\rm rms} = 0.01$ mm. This value of $w_{\rm rms}$ then sets the required sensitivity, $T_{\rm B}^{\rm rms}$, of the radiometers as a function of frequency through equation 28–10. For the VLA we set the $\frac{\lambda}{20}$ requirement at 43 GHz, leading to: $w_{\rm rms} = 0.05$ mm. In its purest form, the inversion of equation 28–13 requires: (i) a sensitive, absolutely calibrated radiometer, (ii) accurate models for the run of temperature and pressure as a function of height in the atmosphere, and (iii) an accurate value for the height of the PWV fluctuations.

Figure 28–14 shows the required sensitivity of the radiometer, $T_{\rm B}^{\rm rms}$, given the benchmark numbers for $w_{\rm rms}$ for the VLA and the MMA and using equation 28–13. It is important to keep in mind that lower numbers on this plot imply that more sensitive radiometry is required in order to measure the benchmark value of $w_{\rm rms}$. The required $T_{\rm B}^{\rm rms}$ values generally increase with increasing frequency due to the increase in A_ν, with a local maximum at the 22 GHz water line, and minima at the strong O_2 lines (59.2 GHz and 118.8 GHz). The strong water line at 183.3 GHz shows a 'double peak' profile, with a local minimum in $T_{\rm B}^{\rm rms}$ at the frequency corresponding to the peak of the line. This behavior is due to the product: $A_\nu \times e^{-\tau_o}$ in equation 28–13. The value of A_ν peaks at the line frequency, but this is off-set by the high total optical depth at the line peak. This effect is most dramatic for the VLA case, where the required $T_{\rm B}^{\rm rms}$ at the 183 GHz line peak is very low.

Carilli, Lay, & Sutton (1998) consider in detail the requirements on the gain stability, sensitivity, and on atmospheric data for absolute radiometric phase correction at the MMA and the VLA. Required sensitivities range from 20 mK at 90 GHz to 1 K at 185 GHz for the MMA, and 120 mK for the VLA at 22 GHz. These target noise values are readily achieved by most planned radiometric phase correction systems. However, gain stability requirements may prove to be a limitation, in particular for uncooled radiometers. The minimum requirement

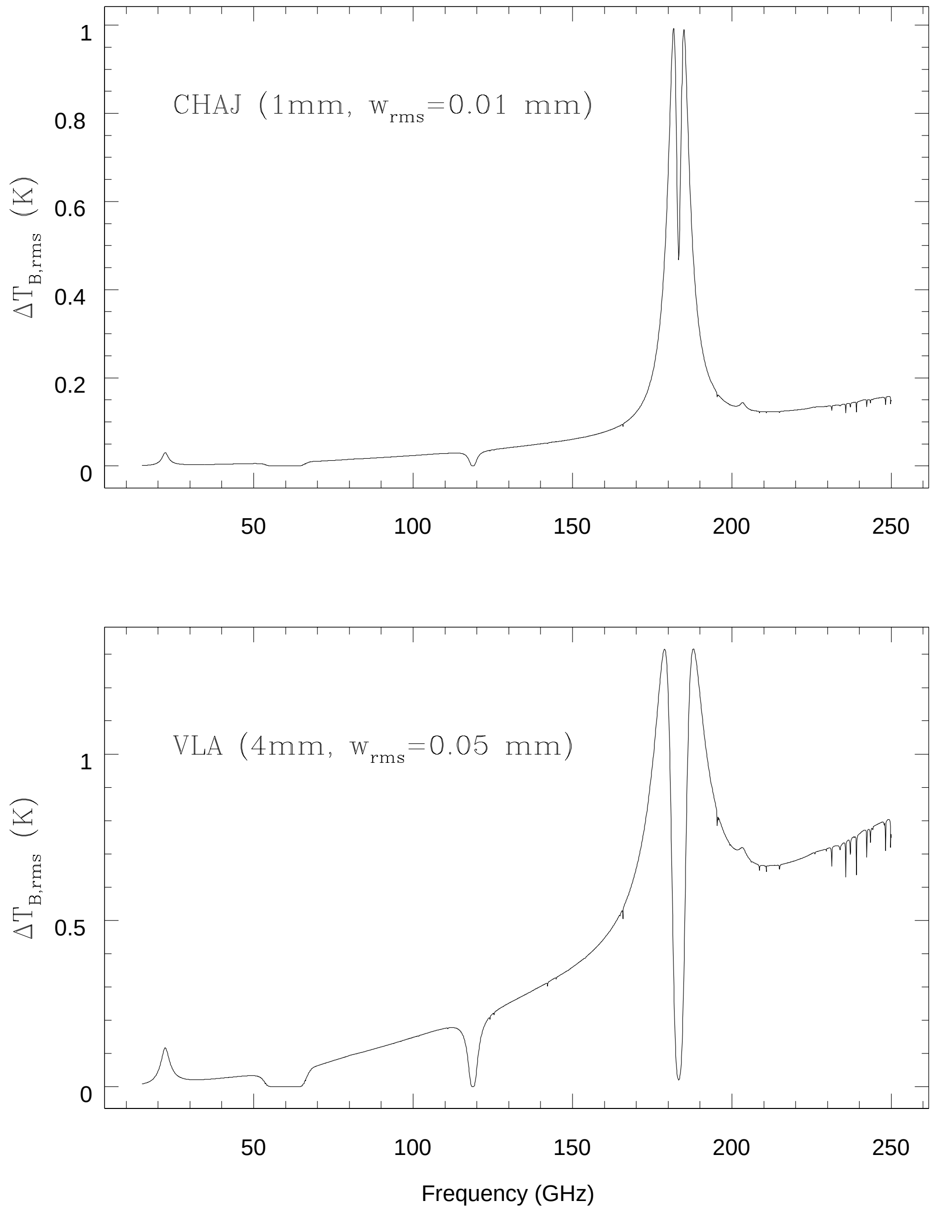


Figure 28–14. The upper frame shows the brightness temperature sensitivities, $T_{\rm B}^{\rm rms}$, required to measure PWV variations to an accuracy of $w_{\rm rms} = 0.01$ mm, corresponding to residual rms phase variations at 230 GHz of 18°, for the MMA site at Chajnantor assuming $w_o = 1$ mm (equation 28–13). The bottom frame shows the corresponding $T_{\rm B}^{\rm rms}$ for the VLA site assuming $w_o = 4$ mm and requiring $w_{\rm rms} = 0.05$ mm, corresponding to residual rms phase variations at 43 GHz of 18°.

is $T_{\rm sys}/T_{\rm B}^{\rm rms} \sim 200$ at 185 GHz at the MMA assuming that the astronomical receivers are used for radiometry. This increases to 2000 for an uncooled system. The stability requirement is 450 for the cooled system at the VLA at 22 GHz.

Converting the measured $T_{\rm B}$ to electrical pathlength requires knowledge of the tropospheric parameters, such as $T_{\rm atm}$, $P_{\rm atm}$, and $h_{\rm turb}$. Carilli et al. (1998) show that to achieve the target $T_{\rm B}^{\rm rms}$ values in Figure 28–14 requires knowledge of the tropospheric parameters to an accuracy of a few percent or better. And even if such accurate measurements are available, fundamental uncertainties in the atmospheric models relating $T_{\rm B}$ and w_o may require empirical calibration of the $T_{\rm B}^{\rm rms} - w_{\rm rms}$ relationship at regular intervals. Lastly, clouds affect the system temperature but not the phase. Separating the effect of clouds (liquid water) from water vapor fluctuations requires a measurement of the power in one of the water vapor lines (22 GHz or 183 GHz).

Carilli et al. (1998) also consider the less demanding technique of making radiometric phase corrections using an empirical calibration of the $T_{\rm B}^{\rm rms}$ – $\phi_{\rm rms}$ relationship to increase the coherence time on source, and perhaps to connect the target source phase to a celestial calibrator. Empirical calibration entails deriving a 'gain factor' relating brightness temperature fluctuation differences between antennas to interferometric phase differences using a measured time series of phases on a bright celestial calibrator.

An example of such a relative calibration is shown in Figure 28–15, using data from the VLA at 22 GHz. Observations were made on the night of October 15, 1998, of the calibrator 0319+415 (3C 84). In the upper frame, the dash line shows the interferometric phase time series measured between antennas 5 and 9, corresponding to a baseline length of about 3 km. The solid line shows the predicted phase time series derived by differencing measurements of the 22 GHz system temperature at each antenna. A single scale factor relating phase fluctuations and fluctuations in system temperature differences was derived from all the data, by requiring a minimum residual rms scatter in the phase fluctuations after applying radiometric phase correction. A constant off-set was also applied to each data set. Note the clear correlation between measured interferometric phase variations, and the phase variations predicted by radiometry. The middle frame shows the residual phase variations after radiometric correction. The rms variations in the raw phase time series before correction are 32°. After applying the radiometric correction, the rms phase variations are reduced to 17°. The lower frame shows the residual phase variations after radiometric correction, but now making a correction for the first and second half of the data separately. The residual rms phase variations are now 13°. The scale factor changes by about 10% over the 36 minutes.

This 'empirically calibrated' radiometric phase correction technique has been implemented successfully at the Owens Valley Radio Observatory and at the IRAM interferometer (Woody & Marvel 1998, Bremer et al. 1997). A number of questions remain to be answered concerning this technique, including: (i) over what time scale and distance will this technique allow for radiometric phase corrections when switching between the source and the calibrator? and (ii) how often will calibration of the $T_{\rm B}^{\rm rms}$ - w_o relationship be required, i.e., how stable are the radiometers and the mean parameters of the atmosphere?

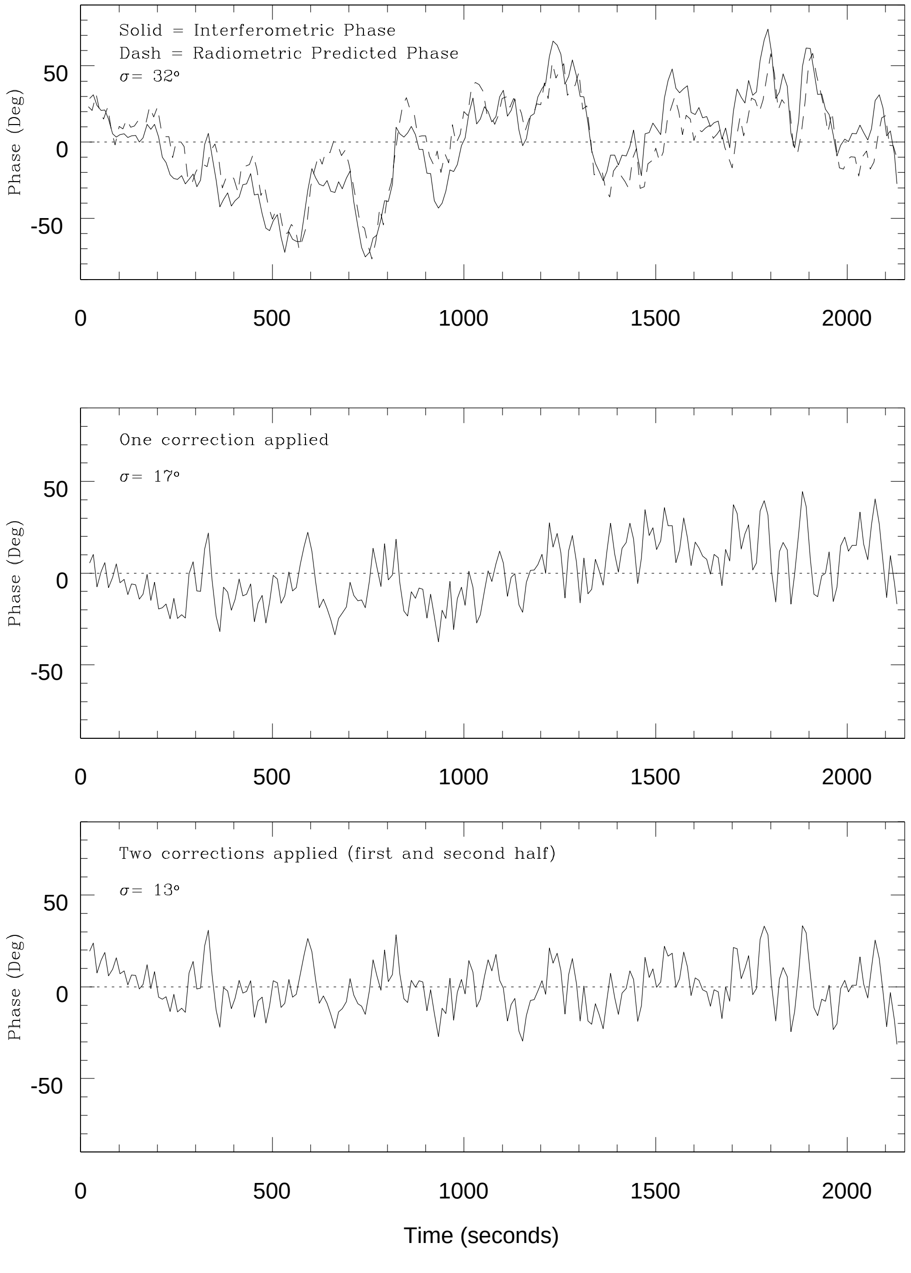


Figure 28–15. Upper frame: The dash line shows the interferometric phase time series at 22 GHz measured between VLA antennas 5 and 9 (baseline length = 3 km). The source observed was the 16 Jy calibrator 0319+415. The solid line shows the predicted phase time series derived by differencing measurements of the 22 GHz system temperature at each antenna. The scale factor relating phase fluctuations and temperature fluctuations was derived from all the data. The middle frame shows the residual phase variations after radiometric phase correction using a single scale factor derived from the entire time series. The lower frame shows the same residuals, but now using corrections derived for the first and second half of the data separately.

Lay (1998) has recently presented an interesting radiometric phase correction method using multifrequency measurements of the 183 GHz water line profile. His method is insensitive to atmospheric parameters, since it relies on using the line profile, and in particular, the 'hinge' points of the lines ≈ half-power points. This method may allow for an absolute radiometric phase correction to be made without great uncertainties due to the atmospheric models.

4. Antennas

Pointing. The planned antennas for the MMA are 10m diameter, as compared to 25m for the VLA. Yet the MMA will operate at frequencies up to 650 GHz, and perhaps as high as 850 GHz – more than an order of magnitude higher frequency than the highest frequencies at the VLA. So while the VLA can operate at cm wavelengths with an rms blind pointing specification of $15''$ ($= \frac{1}{25} \times$ FWHM of the primary beam at 8 GHz), the MMA will require a pointing accuracy to $1''$ or less. Long time scale pointing errors can be caused by thermal gradients (eg. varying insolation) or errors in the antenna pointing model, while short timescale variations can be caused by wind or antenna mechanics (eg. encoder errors or hysteresis).

Pointing errors will cause time dependent 'gain errors', which vary depending on the source position in the primary beam. For example, at 350 GHz the MMA primary beam is approximately Gaussian with FWHM = $18''$. A $3''$ pointing error ($\approx 3\sigma$) would then lead to a 5% reduction in flux of a point source at the pointing center. The effect is even more deleterious for sources close to the half-power point of the primary beam, where the antenna response is a steeply declining function. The amplitude of a source at the half power point would change by about 22% for a 3σ pointing error. Such pointing errors can be the dominant cause of residual errors in a mosaic imaging of sources larger than the primary beam (Cornwell, Holdaway, & Uson 1993).

One method of reducing pointing errors that has been used effectively at 43 GHz at the VLA, and at other mm observatories, is reference pointing (Kestevan 1994, see lecture 3). This technique utilizes pointing scans on celestial calibrators close to the target source to refine the antenna pointing model on timescales of 30 minutes to an hour. This technique will remove long time scale pointing errors, such as slow thermal effects or pointing model errors. This technique reduces the rms pointing errors at the VLA to a few arcseconds under relatively calm weather conditions.

Some observatories are planning active pointing correction techniques. For example, the Green Bank Telescope will employ a laser ranging system referencing the prime focus receiver cabin to the parabolic surface, and to fixed structures on the ground, plus actuators situated on all the telescope panels, to maintain both absolute shape of the off-set parabola, and absolute pointing position. Corrections can then be made on timescales of seconds.

Aperture Efficiency. Given a required aperture efficiency, η, the Ruze formula implies that the required surface rms accuracy, $\sigma_{\rm rms}$, must decrease linearly with wavelength as: $\sigma_{\rm rms} = (-\ln \eta)^{1/2} \frac{\lambda}{4\pi}$ (Lecture 3). At 230 GHz the required accuracy of the antenna panels to achieve $\eta = 50\%$ is then 90μm. The panels must also be aligned into the proper parabolic shape with comparable accuracy.

One powerful method that has been developed in the last few years to determine for errors in antenna shape is interferometric holography (Kestevan 1993, see Lecture 3). Surface holography relies on the Fourier transform relationship between the complex electric field distribution across the aperture and the far field complex voltage pattern of the antenna. The technique measures the complex voltage pattern of the primary beam using a two dimensional raster scan observation of a celestial source with one antenna, with the second 'reference' antenna tracking the source at the pointing center. The reference antenna defines the unit amplitude and zero phase measurement for the pointing center. The Fourier transform of the visibilities gives an image of the complex electric field distribution across the antenna surface, the phase of which dictates the surface errors. The VLA aperture efficiency improved by a factor of about three at 43 GHz (from 15% to 40%) after panel adjustments were made using holographic measurements as a guide.

Baseline Errors. The phase errors introduced due to errors in the positions of the telescopes are inversely proportional to the observing wavelength:

$$\Delta\phi = \frac{2\pi}{\lambda} \times \Delta b \times \Delta\theta \tag{28–14}$$

where $\Delta\theta$ is the angular separation of the source and calibrator, and Δb is the error in the baseline length determination. Typical calibrator-source distances for the MMA will be of order 10°. Hence, to keep $\Delta\phi$ due to baseline errors below 10° at 230 GHz will require baseline determinations accurate to 0.2 mm.

5. Electronics in Millimeter Interferometry

A fundamental difference between cm and mm receivers arises due to the fact that it is currently not possible to build low noise transistor amplifiers (LNAs) that operate above 115 GHz (Maas 1992), so the first element in the receiver must be a mixer. Since the LNA is the first element in cm wavelength receivers, the noise temperature for the receiver is determined essentially by the LNA itself, which is usually based on a low noise HEMT (High Electron Mobility Transistor) amplifier with a receiver temperature of about 10 K and a gain of about 30 dB.

At mm wavelengths the first element must be a mixer. Conventional Schottky-barrier diode mixers can be used up to 230 GHz, but these systems can have significant losses, G $\approx$ 0.2, meaning the mixer attenuates the incoming signal before amplification. Using equation 28–3, the effective receiver noise temperature referenced to the unamplified sky signal becomes: $T_{\rm rec} = T_{\rm mix} + 5{\times}T_{\rm IF}$, where $T_{\rm IF}$ is the IF amplifier noise temperature. This means that even if the noise contribution from the mixer ($T_{\rm mix}$) is low, the loss in the mixer has increased the effective noise temperature of the amplifier by a factor of five relative to a normal heterodyne receiver.

The solution has been to use SIS (superconductor-insulator-superconductor) junctions operating at liquid Helium temperatures ($\leq$ 4 K) for the first mixers in mm interferometers. These devices employ photon assisted quantum mechanical tunneling to achieve low noise temperatures ($T_{\rm mix} \approx 10$ K), and high gains ($\rm G_{mix} \approx 1$), with instantaneous bandwidths up to a few GHz.

Millimeter receivers also employ quasi-optical techniques for the LO signal injection, in which the LO signal is injected directly into the feed along with the sky signal using a polarizing reflector. The LO and signal then propagate through free space to the mixer.

5.1. The Quantum Limit

What is the limit to heterodyne receiver technology, and can can we extrapolate the technique to much higher frequency, and possibly build heterodyne receivers for optical interferometry? The fundamental limit is set by quantum mechanics (Bester et al. 1990, Townes et al. 1998, Thompson, Moran, & Swenson 1986; see Lecture 33).

The Heisenberg uncertainty principle for the non-commuting variables of energy and time, states that: $\Delta E \times \Delta t = \frac{h}{2\pi}$, where h is Planck's constant. Hence to localize an 'object' or 'quantity' in time, Δt, necessitates an increase in the uncertainty in the energy, ΔE (any individual measurement of a system perturbs the state of the system). Conversely, it takes increasingly long to make an accurate measurement of the energy of a system.[5] The uncertainty in the number of photons is given by: $\Delta n_{\rm ph} = \frac{\Delta E}{h\nu}$, while the uncertainty in the phase of the photons is given by: $\Delta\phi = 2\pi\nu\Delta t$. The uncertainty principle becomes:

$$\Delta\phi \times \Delta n_{\rm ph} = 1\ s^{-1}{\rm Hz}^{-1}$$

Electronic devices such as mixers or amplifiers that conserve phase for the output, i.e., $\Delta\phi \leq 1$ rad, must pay a quantum mechanical 'measurement price' in terms of the uncertainty in the number of photons: $\Delta n \geq 1$ photon per mode, where a mode $\equiv$ per Hz per second (Lecture 33). Since a phase-conserving receiver must localize the signal in time, Δt, it must pay the measurement price in terms of noise temperature: $kT_{\rm sys}^{\rm QM} = \Delta E$. The quantum limit to the system noise then becomes:[6]

$$T_{\rm sys}^{\rm QM} = \frac{h\nu}{k\Delta\phi} \geq \frac{h\nu}{k} \quad \text{for } \Delta\phi \ \leq \ 1 \text{ rad} \tag{28–15}$$

The implied absolute minimum noise temperature at 1.4 GHz is: $T_{\rm sys}^{\rm QM}$ = 0.07 K, while at 230 GHz: $T_{\rm sys}^{\rm QM}$ = 11 K. The quantum limit to the noise temperature in the optical, where $\nu \approx 6{\rm x}10^{14}$ Hz^{-1}, is: $T_{\rm sys}^{\rm QM}$ = 30,000 K. For

[5] From the point of view of signal processing, Radhakrishnan (1998) points out that the uncertainty principle can be viewed as the product of the total power in the signal, which is linearly proportional to the bandwidth, $\Delta E \propto \Delta\nu$, and the required sampling rate, which is dictated by the Nyquist rate and hence is inversely proportional to the bandwidth, $\Delta t \propto \Delta\nu^{-1}$.

[6] This excess noise arises from the standard relationship between stimulated and spontaneous emission. The amplified signal from a phase conserving amplifier is analogous to 'stimulated emission', which then requires a corresponding amount of spontaneous emission, as dictated by the Einstein A and B coefficients: $A_{ij} = \frac{h\nu^3}{c^2} B_{ij}$.

comparison, direct detection devices such as optical CCDs do not conserve phase, thereby allowing $\Delta\phi \rightarrow \infty$ such that $\Delta E \rightarrow 0$, or $\Delta n \rightarrow 0$. Such photon counting devices are typically source photon noise limited.

At radio frequencies the noise contribution from background sources, such as the cosmic microwave background at 3 K, is comparable to, or larger than, the quantum noise limit in a heterodyne receiver. Hence the additional 'quantum noise' for phase conserving electronics is not a dominant factor is determining the noise characteristics for radio receivers, thereby mitigating the advantage of direct detectors.

At optical frequencies, the 3 K background and other background contributions are completely swamped by the quantum noise. While it is possible to build optical mixers and heterodyne receivers, eg. using laser LO signals, it is clearly impractical in astronomy where the sky signals are weak and where very low noise direct detection devices are available. The radical noise reduction for direct detectors in the optical relative to quantum limited detectors makes it advantageous to do interferometry using optics, and simply measure the resulting interference pattern directly with photon counting devices.

As a practical example, consider the powerful radio galaxy Cygnus A. The distance to Cygnus A is 230 Mpc and the optical and radio luminosities are both about 10^{45} erg s^{-1} (note that the optical emission is mostly starlight, while the radio emission is synchrotron radiation from the jet-powered radio lobes). At 100 GHz the flux density is about 10 Jy = 10^{-22} erg cm^{-2} s^{-1} Hz^{-1}, while the optical flux density is about 10 mJy = 10^{-25} erg cm^{-2} s^{-1} Hz^{-1}. Assuming one uses a 10 m telescope for both the optical and radio measurements, and dividing by $h\nu$, gives the number of source photons per mode coming to the detector. In the radio this value is $n_{\rm src}^{\rm radio} = 0.1$ s^{-1} Hz^{-1}, and in the optical the value is $n_{\rm src}^{\rm opt} = 10^{-8}$ s^{-1} Hz^{-1}. It can be shown that the SNR for a quantum limited heterodyne receiver behaves as $\mathrm{SNR_{het}} = n_{\rm src} \times (\Delta\nu t)^{\frac{1}{2}}$, where t is the integration time and $\Delta\nu$ is the bandwidth. Typical bandwidths are about 10^9 Hz in the radio and 10^{14} Hz in the optical. For Cygnus A in the radio the $\mathrm{SNR_{het}^{radio}} = 3000$, while in the optical the $\mathrm{SNR_{het}^{opt}} = 0.1$ per second of integration time. For a photon counting device one can show that the $\mathrm{SNR_{DD}} = (n_{\rm src}\Delta\nu t)^{\frac{1}{2}}$, or for Cygnus A the optical $\mathrm{SNR_{DD}^{opt}} = 1000$ per second. To achieve an optical SNR of 1000 using a quantum limited heterodyne receiver would take 10^8 seconds for Cygnus A, as opposed to 1 second for the direct detection device.

One advantage of phase conserving electronics is that, once the signal is amplified coherently, it can be split many times without decreasing the SNR. This is not the case for interferometery using optics plus direct detectors, where the unamplified signal must be split, and hence the SNR decreases linearly with the number of elements (Townes et al. 1998). Other advantages of narrow bandwidth heterodyne systems include simpler delay lines (see Equation 2–16), and the reduced effect of 'seeing' on visibility amplitude calibration. Such techniques are being explored in the near-infrared by a number of groups (Bester et al. 1990, 1994, Townes et al. 1998).

Acknowledgments. We thank K. Desai, G. Blake, D. Hughes, and J. Lim for allowing us to use figures from previous publications. We thank Anantharamiah, Radhakrishnan, and V. Dhawan for lively discussions on the quantum limit.

References

Asaki, Y., Saito, M., Kawabe, R., Morita, K., Sasao, T. 1996, *Radio Science*, 31, 1615

Bagri, D. S. 1994, VLA Test Memo. No. 184

Bagri, D. S. & Lillie, P. 1993, VLA Test Memo. No. 170

Barrett, A. H. & Chung, V. K. 1962, *J. Geophys. Res.*, 67, 4259

Barvainis, R. 1998, in *Highly Redshifted Radio Lines*, eds. Carilli, Radford, Menten, & Langston (San Francisco: PASP)

Bean, B. R. & Dutton, E. J. 1968, *Radio Meteorology*, (New York: Dover), p. 7

Beckwidth, S. V. & Sargent, A. I. 1996, Nature, 383, 139

Bester, M., Danchi, W. C., & Townes, C. H. 1990 in *SPIE 237: Amplitude and Intensity Spatial Interferometry*, ed. J. Breckinridge, (Washington: SPIE), p. 40

Bester, M., Danchi, W. C., & Townes, C.H. 1994 in *SPIE 2200: Amplitude and Intensity Spatial Interferometry II*, ed. J. Breckinridge, (Washington: SPIE), p. 274

Blake, G. A., Sutton, E. C., Masson, C. R., & Phillips, T. G. 1987, *ApJ*, 315, 621

Blain, A. W., Ivison, R. J., & Smail, I. 1998, *MNRAS*, 296, 29p

Bremer, M., Guilloteau, S., & Lucas, R. 1997, in *Science with Large Millimetre Arrays*, eds. P. Shaver (ESO: Garching)

Brown, R. L. 1996, in *Cold Gas at High Redshift*, eds. Bremer, van der Werf, Rottgering, & Carilli, (Dordrecht: Kluwer)

Butler, B. J. 1996, VLA Scientific Memo. No. 170

Butler, B. J. 1997, MMA Memo. No. 188

Butler, B. J. 1998, VLA Test Memo. No. 212

Carilli, C. L., Lay, O. P., & Sutton, E. C. 1998, MMA Memo. No. 210

Carilli, C. L., Holdaway, M. A., & Sowinski, K. 1996, VLA Scientific Memo. No. 169

Carilli, C. L. & Holdaway, M. A. 1996, VLA Scientific Memo. No. 171

Carilli, C. L. & Holdaway, M. A. 1997, MMA Memo. No. 173

Clark, B. G. 1973a, VLA Computer Memo. No. 103

Clark, B. G. 1973b, VLA Computer Memo. No. 105

Clark, B. G. 1974, VLA Computer Memo. No. 112

Clark, B. G. 1987, VLA Computer Memo. No. 158

Cornwell, T. J., Holdaway, M. A., & Uson, J. M. 1993, *A&A*, 271, 697

Coulman, C. E. 1990, in *Radio Astronomical Seeing*, eds. J. Baldwin & S. Wang, (Pergamon: New York), p. 11.

Counselman, C. C. et al. 1974, *Phys. Rev. Lett.*, 33, 1621

Dicke, R. H., Beringer, R., Kyhl, R. L., & Vane, A. B. 1946, *Physical Review*, 70, 340

van Dishoeck, Ewine, Blake, G. A., Draine, B. T., & Lunine, J.I. 1993, in *Protostars and Planets III*, eds. Levy & Lunine (Tucson: Univ. Arizona Press)

Drashkikh, A. F. & Finkelstein, A. M. 1979, *Astrophy. Space Sci.*, 60, 251

Garratt, J. R. 1992, *The Atmospheric Boundary Layer*, (Cambridge Univ. Press: Cambridge), p. 11.

Hinder, R. & Ryle, M. 1971, *MNRAS*, 154, 229

Hogg, D. C., Guiraud, F. O., & Decker, M. T. 1981, *A&A*, 95, 304

Holdaway, M. A. & Woody, D. 1998, MMA Memo. No. 223

Holdaway, M. A. 1998, MMA Memo. No. 221

Holdaway, M. A. 1997, MMA Memo. No. 186

Holdaway, M. A., Radford, S., Owen, F. N., & Foster, S. 1995, MMA Memo. No. 139

Holdaway, M. A. & Owen, F. N. 1995, MMA Memo. No. 126

Holdaway, M. A. Owen, F. N., & Rupen, M. P. 1994, MMA Memo. No. 123

Holdaway, M. A. 1992, MMA Memo. No. 84

Holdaway, M. A. & Pardo, J. R. 1997, MMA Memo. No. 187

Hughes, D. & Dunlop, J. 1998, in *Highly Redshifted Radio Lines*, eds. Carilli, Radford, Menten, & Langston (San Francisco: PASP)

Hughes, D. et al. 1998, *Nature*, 394, 241

Kestevan, M. 1994, VLA Test Memo. No. 183

Kestevan, M. 1993, VLA Test Memo. No. 169

Kogan, L. 1997, *ELINT*, Astronimical Image Processing System

Kutner, M. L. & Ulich, B. L. 1981, *ApJ*, 250, 341

Lay, O. P. 1998, MMA Memo. 209

Lay, O. P. 1997a, *A&AS*, 122, 547

Lay, O. P. 1997b, *A&AS*, 122, 535

Liebe, H. J. 1989, International Journal of Infrared and Millimeter Waves, 10, 631

Lim, J., Carilli, C. L., White, S., Beasley, A., & Marson, R. 1998, *Nature*, 392, 575

Levy, E. H. & Lunine, E. H. 1993, *Protostars and Planets III*, (Tucson: Univ. Arizona Press)

Maas, S. A. 1992, *Microwave Mixers*, (Artech House: Boston)

Madau, P., Ferguson, H. C., Diskinson, M. E., Giavalisco, M., Steidel, C. C., & Fruchter, A. 1996, *MNRAS*, 283, 1388

Mehringer, D. M., Snyder, L. E., Miao, Y., & Lovas, F. J. 1997, *ApJ*, 480, 71

Menten, K. M., Carilli, C. L., & Reid, M. J. 1998, in in *Highly Redshifted Radio Lines*, eds. Carilli, Radford, Menten, & Langston (San Francisco: PASP)

Poynter & Pickett 1985, Applied Optics 24, 2235 http://spec.jpl.nasa.gov/

Rosenkranz, P. 1989, in *Atmospheric Remote Sensing by Microwave Radiometry*, ed. M. Janssen, (New York: Wiley and Sons)

Schilke, P., Groesbeck, T. D., Blake, G. A., & Phillips, T. G. 1997, *ApJS*, 108, 301

Staelin, D. H. 1966, *J. Geophys. Res.*, 71, 2875

Staguhn, J., Harris, A., Plambeck, R., & Welch, W. J. 1998, in *SPIE 3357: Advanced Technology MMW, Radio, and Terahertz Telescopes*, ed. T. Phillips, (Washington: SPIE), p. 432

Sramek, R. 1990, in *Radio Astronomical Seeing*, eds. J. Baldwin & S. Wang, (Pergamon: New York), p. 21

Sutton, E. C. & Hueckstaedt, R. M. 1997, *A&AS*, 119, 559

Tatarskii, V. I. 1978, Wave Propagation in Turbulent Media, (New York: Wiley)

Taylor, G. I. 1938, *Proc. R. Soc. London*, A 164, 476

Thompson, A. R., Moran, J. M., & Swenson, G. W. 1986, *Interferometry and Synthesis in Radio Astronomy*, (Wiley: New York)

Townes, C. H. et al. 1998 in *SPIE 3350: Astronomical Interferometry*, ed. R.D. Reasonberg, (Washington: SPIE), p. 908

Waters, J. W. 1976, in *Methods of Experimental Physics, Vol 12, Astrophysics*, Part B: Radio Telescopes, (Academic Press:New York), Chaps. 2 and 3

Welch, W. J. 1994, in *Astronomy with Millimeter and Submillimeter Wave Interferometry*, eds. M. Ishiguro & W.J. Welch, (San Francisco: PASP), p. 1

Westwater, E. R. & Guiraud, F. O. 1980, *Radio Science*, 15, 947

Wilner, D. J., Ho, P. T. P., & Rodriguez, L. F. 1996, *ApJ*, 469, 216

Wiklind, T. & Combes, F. 1998, in *Highly Redshifted Radio Lines*, eds. Carilli, Radford, Menten, & Langston (San Francisco: PASP)

Woody, D. & Marvel, K. 1998, in *SPIE 3357: Advanced Technology MMW, Radio, and Terahertz Telescopes*, ed. T. Phillips, (Washington: SPIE), p. 442

Wright, M. C. H. 1996, *PASP*, 108, 520

Yun, M. S., Mangum, J., Bastian, T., Holdaway, M., & Welch, J. 1998, MMA Memo. No. 211

Zheng, Y. 1985, Ph.D. Thesis, Iowa State University.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

29. Long Wavelength Interferometry

W. C. Erickson
Astronomy Department, University of Maryland, MD 20742, U.S.A. and School of Mathematics and Physics, University of Tasmania, Hobart, Tasmania, 7001, Australia

Abstract.
A long wavelength capability has recently been implemented at the VLA with the installation of instrumentation for operation at 4 m wavelength (74 MHz). The first A-configuration observations were made with this new system in February, 1998. This chapter will discuss the capabilities of this system, the types of programs that it can undertake, and its limitations and problems.

1. Introduction

First, I should define what I mean by "long wavelength" since it can mean anything from 1 mm to 1000 m; I'm defining it to be about 1 m or longer ($f \leq$ 300 MHz), and especially the 4 m band at the VLA.

Long wavelength work has its own joys and frustrations. The long wavelength region is the least explored portion of the electromagnetic spectrum. Its greatest attraction is that it is still possible to carry out forefront research with relatively simple instruments and programs; some useful long wavelength instruments cost thousands rather than hundreds of thousands or millions of dollars, and many observing programs are simple enough that large group efforts are not required. Some stalwart individuals find this ability to "do your own thing" quite captivating. It must be recognized, however, that this situation will change over the next decade as long wavelength instruments become more sophisticated. The VLA is one of the leaders in this direction, and it now possesses unique capabilities which can be utilized for highly original research at long wavelengths. These capabilities include unprecedented angular resolution of ~25 arcsec resolution with 10 to 20 mJy/beam sensitivity. Traditionally, the VLA has operated at wavelengths well away from the edges of the radio band transmitted by the Earth's atmosphere. This situation is changing as new systems are implemented for millimeter wavelengths and for meter wavelengths. These systems provide the observer with many new opportunities but they also confront the observer with many new problems.

Many similarities exist between the problems encountered at mm wavelengths and those encountered at meter wavelengths. Firstly, sensitivities tend to be lower than in other bands. At mm wavelengths this is because collecting areas are low and system temperatures relatively high. At meter wavelengths high Galactic background temperatures result in very high system temperatures and only narrow interference-free bandwidths are available. Secondly, rapid atmospheric phase fluctuations imply short coherent integration times. At mm wavelengths these fluctuations are tropospheric while at meter wavelengths they are ionospheric. Thirdly, because of the rapid phase fluctuations and a paucity of good calibration sources, calibration tends to be difficult in both bands. Finally, atmospheric conditions, either tropospheric or ionospheric, often destroy observations completely. Some type of dynamic telescope scheduling which takes atmospheric conditions into consideration is highly advisable. Of course, great differences also exist between the bands. At meter wavelengths antenna pointing and surface accuracy is no problem. Receivers are cheap and simple; interference, both natural (lightning, dust static, solar bursts, etc.) and man-made (digital equipment, arcing contacts, radio transmitters, etc.), is often a severe problem; and radio source confusion, rather than thermal noise, often limits a system's sensitivity. Considering these difficulties, one might ask, "Why bother with long wavelength work?". In this discussion, I will attempt to answer this question.

2. Features of the Radio Sky at Long Wavelengths

2.1. The Galactic Background

The Galactic background is the dominant feature of the long wavelength radio sky and, as we will see, has important implications upon various aspects of interferometric work. Most importantly, the background emission dominates the system noise in this wavelength range. At 4 m wavelength the data summarized by Cane (1978) yield an average brightness temperature of the Galactic Polar regions of 1700 ± 70 K. This represents the minimum system temperature that can be obtained. At meter wavelengths the background temperature near the Galactic Plane is roughly an order of magnitude higher than at the Poles, and at decameter wavelengths ($f \sim 30$ MHz) distributed HII along lines-of-sight in the Plane causes an absorption trough to develop. (Note: the physical temperature of HII - $\sim$8000 K - is much lower than the background temperature near 30 MHz so HII regions appear in absorption.) At even longer wavelengths the absorption trough along the Plane broadens and deepens until the polar regions are brighter than the Plane at hectometric wavelengths ($f \sim 3$ MHz). Figure 1 is a representative map of the background distribution at 150 MHz. The brightness temperatures of most of the features on this map scale in frequency as $T_b \sim f^{-2.55}$ which means that for 74 MHz the brightness temperatures of features on this map should be multiplied by four.

It is easy to construct simple transistor preamplifiers with noise temperatures of about 100 K in the 4 m band, so the antenna temperature caused by the background obviously dominates the receiver noise. Complex, low-noise preamplifiers are not needed for this work. It must also be recognized that the system temperature will be a strong function of the pointing direction of the antenna. This can sometimes lead to dynamic range or calibration problems.

The Galactic background radiation is synchrotron emission from cosmic ray electrons possessing energies of several hundred MeV as they spiral in Galactic magnetic fields. Its study represents one of the very few ways in which the distribution of cosmic rays throughout the Galaxy can be explored. In particular, small, dense HII regions can be used as opaque walls to shield the distant background emission, permitting a study of the foreground cosmic ray emissivity along various lines-of-sight towards these HII regions. Such three dimensional studies of the cosmic ray emissivity are of obvious importance to theories of cosmic ray origin and propagation.

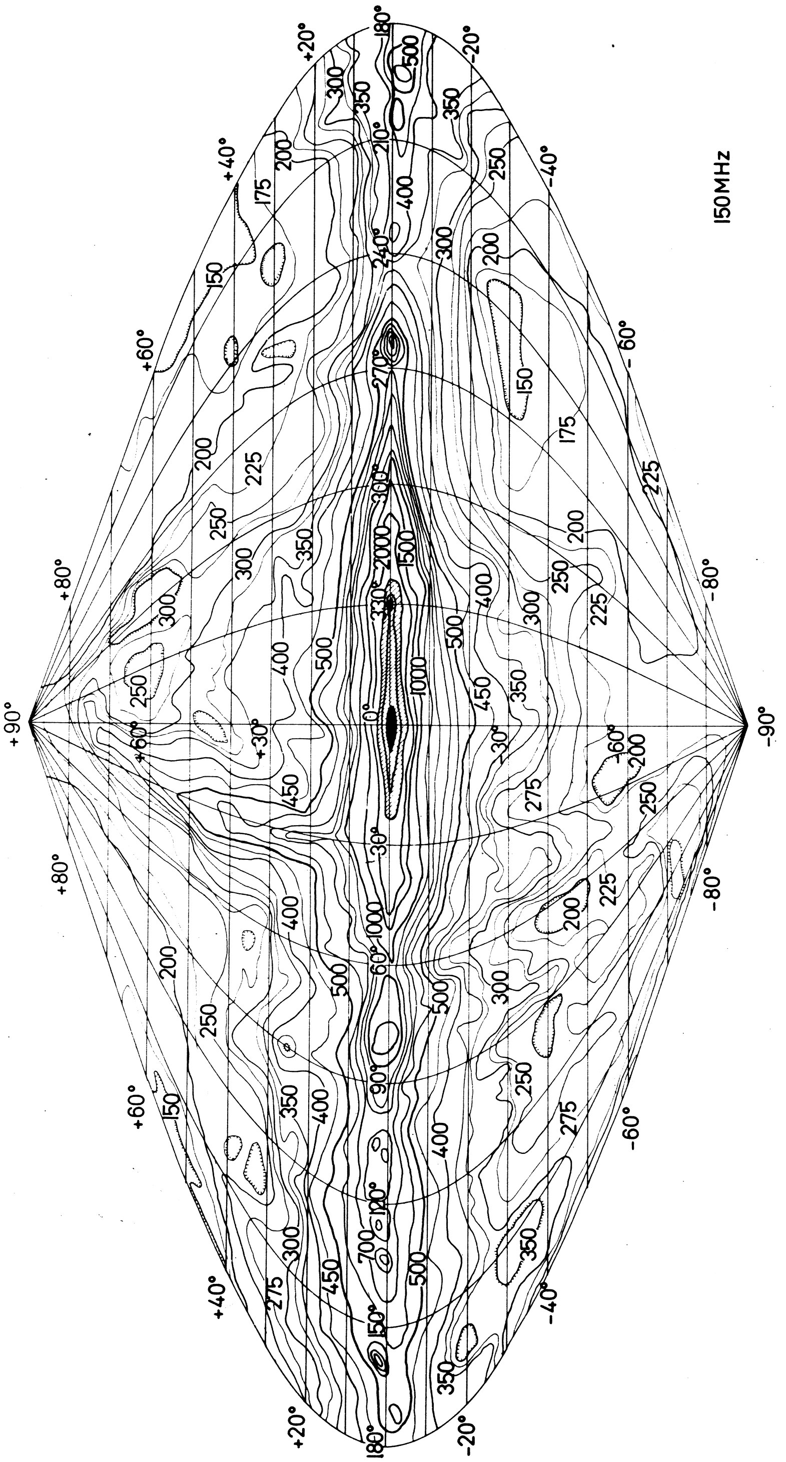


Figure 29–1. An all-sky map of the 150 MHz brightness temperatures in Galactic coordinates. The data were obtained from surveys at 85, 150, and 178 MHz. (Landecker & Wielebinski 1970)

2.2. Radio Sources

With flux densities of ~22,000 Jy and ~18,000 Jy, respectively, the radio sources Cas-A and Cyg-A are, by far, the strongest discrete sources in the radio sky at 4 m wavelength. The next strongest sources; Tau-A, Vir-A, and the undisturbed Sun, are at the 1000 Jy level. However, radio outbursts from the Sun can reach 10^9 Jy for short periods of time, dominating all other sources and wiping out most observing programs. After the twenty or so strongest sources there are a myriad of weaker ones and one can expect to find a considerable number of sources in every field-of-view. Usually there will be adequate sources within a field for calibration, but elimination of the sidelobes from the numerous sources, both within the field and outside of it, can be difficult.

The bulk of the source radiation observed at long wavelengths is generated by the incoherent synchrotron process; emission from high energy electrons as they spiral magnetic field lines. This emission is broad-band, primarily because the energy spectra of the electrons is broad, and the apparent brightness distributions of sources generally change only slowly and rather subtlety with wavelength. It is usually appropriate to carry out detailed studies of source structure at shorter wavelengths, where instrumental resolution and sensitivity are higher, and then to combine these observations with long wavelength ones for spectral index studies of the various source components. The long wavelength observations provide a long "lever arm" for accurate spectral index determinations. Such spectral studies are important, for example, to theories of cosmic ray origin and particle acceleration in Galactic supernova remnants, and to constraining self-absorption and spectral aging effects in extragalactic radio sources.

The electrons that emit long wavelength radiation have long synchrotron lifetimes, often exceeding the Hubble age. At long wavelengths we can observe the source population of electrons which has not been affected by synchrotron losses (except at very high red shifts where the short wavelength portion of the spectrum may be shifted to a long wavelength). Radio galaxies generally have low-brightness, steep spectrum emitting regions in which radiation lifetimes are long, thereby offering a glimpse at source activity long ago. Such regions usually are found in bridges between double sources and in radio source halos.

The spectra of most compact radio sources and of the hot spots in the lobes of some double sources turn down at long wavelengths. This could be caused by a variety of absorption or radiation suppressing mechanisms such as synchrotron self-absorption, Razin-Tsytovich suppression, absorption by thermal gas, or a low-energy cutoff in the source's electron energy spectrum. All of these mechanisms exhibit different spectral signatures and an understanding of them will add considerably to our knowledge of the physics of these sources. At present, cutoffs in the electron energy spectra are the generally preferred mechanism.

Objects at very high redshifts tend to have steep spectra at long wavelengths as the steep short wavelength portions of their spectra are shifted into the long wavelength range. Steep long wavelength spectra have been used by Blundell et al. (1998) as a criterion to discover high redshift objects. An important new class of ultra-steep-spectrum quasars has also been discovered by De Breuck et al. (1998).

In addition to incoherent synchrotron emission, coherent emission processes occur primarily at long wavelengths. As is well known from the Larmor formula, the total power radiated by a charge, q, undergoing an acceleration, a, is proportional to q^2a^2. If electrons are moving incoherently, q is just the electronic charge but if a cloud of electrons with a charge density, N, move coherently, then the effective radiating charge is the total charge within a volume of dimension $(\lambda/2\pi)^3$, i.e. $q \sim N(\lambda/2\pi)^3$, so the total radiation goes as $\sim \lambda^6$. Thus coherent emission processes become far more important as the wavelength is increased. Prominent examples include pulsar emission, solar and stellar radio bursts, and Jovian bursts. Coherent emission is distinguished by rapid time and spectral variations. It often leads to a wealth of information concerning plasma densities and plasma flows in the source region and/or to models of the source's electrodynamics.

2.3. Propagation Effects

Most propagation effects in plasmas; such as thermal absorption, refraction, and Faraday rotation, scale as λ^2 and are important at longer wavelengths. This allows many interesting studies of the medium between the source and the observer but it is a mixed blessing. It is sometimes difficult to distinguish between intrinsic source properties and propagation effects; interesting source properties are often obscured by propagation effects and the ionospheric plasma disturbs long wavelength observations in many ways.

The Interstellar Medium. The interstellar medium does not, in general, cause appreciable absorption at 4 m although strong absorption could occur in the vicinity of compact extragalactic sources, and thermal absorption certainly does occur in dense HII regions. The line-of-sight towards Sgr-A East in the Galactic Center, for example, is completely obscured by absorption at this wavelength and many other lines-of-sight passing near to the Center will also suffer absorption. However, the absorption is patchy and relatively transparent regions are found within 15′ of Sgr-A East. Similar absorption regions occur in external galaxies.

Diffractive interstellar scintillations do not occur except for sources with μarcsecond dimensions, such as pulsars. However, diffractive temporal broadening may smear the time profiles of some pulsars and make them into continuum sources (Cordes 1990). Interstellar scattering produces angular broadening of source sizes that is typically about $0.1''$ at $\lambda = 4\ m$ for lines-of-sight out of the Plane (i.e. $|b| > 10°$), generally much smaller than the $25''$ VLA beamwidth. A typical scattering size for a path along the Plane with $|l| > 40°$would be $2''$, and towards the Galactic Center ($|l| < 40°$) the scattering can exceed the $25''$ beamwidth (Fey, Spangler, & Cordes 1991). There also exist localized regions of greatly enhanced scattering, e.g. the Cygnus region at $l = 80°$.

Interstellar Faraday rotation is large, perhaps 1 to 10,000 radians at 4 m, and will defeat most attempts to measure linear polarization in this band. The only possible sources are solar system objects or nearby pulsars.

The Interplanetary Medium. The interplanetary medium does not generate appreciable absorption at 4 m but does produce amplitude scintillations for arcsecond sources along lines-of-sight within 30° of the Sun (Scott, Coles, & Bourgois 1983). Using a 4 m wavelength, these phenomena were first studied by Hewish, Scott, & Wills (1964) and also led to the discovery of pulsars (Hewish et al. 1968). Closer to the Sun, within about 15°, coronal scattering will broaden the apparent angular sizes of all sources to several arc-minutes (Erickson 1964). These scattering and scintillation phenomena are known to be time variable and strongly dependent upon solar activity.

3. Ionospheric Effects

3.1. Ionospheric Absorption, Scintillation, and Faraday Rotation

Ionospheric absorption is not an appreciable problem in the 4 m band. Measurements indicate that it should only be about 0.1 dB in the daytime and 0.01 dB at night (Lawrence, Little, & Chivers 1964). However, strong absorption of 2-3 dB could occur during intense ionospheric disturbances following large solar flares. These absorption events generally last for approximately 10 minutes.

Amplitude scintillation occurs when multipath rays deflected by the ionosphere converge on the telescope to produce constructive and destructive interference effects. The deviations at 4 m are usually small enough that $\sim$ 100% scintillations do not occur at the VLA site. However, in the polar and equatorial regions of the Earth strong ionospheric scintillations are common. Smaller effects at the 2% to 20% level can be anticipated at the VLA.

Time variable ionospheric Faraday rotation often amounts to several turns at 4 m wavelength (Garriott, Smith, & Yuen 1965). However, the interstellar effects are already so large that this added complication is probably of little importance. Correction for ionospheric Faraday rotation is important for polarization work in the 1 m and 20 cm bands. We have recently developed procedures for making these corrections (Erickson et al. 1998).

3.2. Ionospheric Phase Fluctuations

The phase fluctuations caused by trans-ionospheric propagation are the most serious problem in 4 m observations. Over long, A-configuration, baselines phase winding of 1 radian/minute or more can be expected, especially near sunrise and sunset when ionospheric densities change rapidly. Even faster phase winding may occur when the ionosphere is disturbed, but there also are quiet periods when the winding is an order-of-magnitude less. A 10 second integration time is the maximum that can be used to follow the phase winding in a reasonable fashion. A shorter integration time would be preferable but the VLA on-line system cannot handle the data flow resulting from shorter integrations in spectral line mode and this mode is required to excise certain 100 kHz interference spikes that are generated by the system. In the C and D-configurations, the phase winding will be slower and longer integrations may be used.

A dual-frequency technique has been found useful to slow down the phase winding. Observations are made simultaneously at 1 m and at 4 m (the "4P" observing mode). The 1 m data are calibrated in the standard fashion using self-calibration techniques and the antenna-based phases are determined. The 1 m phase winding is four times slower than that at 4 m, so longer integrations are feasible at this wavelength. The higher sensitivity and smaller primary beamwidth ($\sim 2.5°$) at 1 m also provide more detectable, unconfused sources for possible use as calibrators. The antenna phases at 1 m are then multiplied by the wavelength ratio and applied to the 4 m data. As shown in Figure 2,

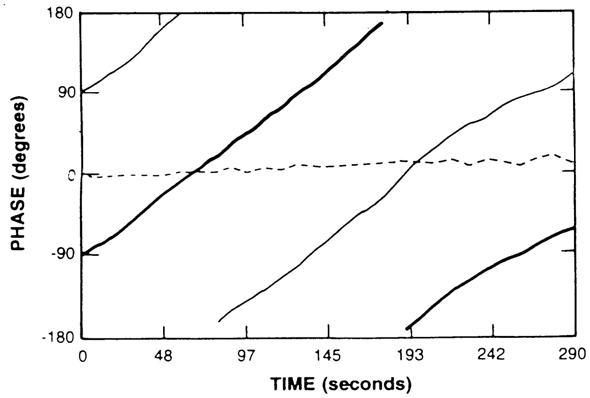


Figure 29–2. Plot of visibility phase (in degrees) vs time (in seconds) on a ~ 15 km baseline for a strong point source (Kassim et al. 1993). The thick solid line is the 74 MHz phase plotted at an arbitrary starting point of -90°. The ~ 360°phase wind is due to the ionosphere. The thin solid line is the 330 MHz phase after multiplication by the frequency ratio (330/74) and started at an arbitrary phase of +90°. The dashed line represents the difference between the two and illustrates how well the "scaled" 330 MHz phase tracks the ionospheric phase variations at 74 MHz.

this effectively stops the 4 m phase winding. However, since the 1 m phases are multiplied by the wavelength ratio, phase noise at 1 m is amplified by a factor of four and the maximum per baseline phase noise that can be tolerated is about 0.1 radians. This leads to a requirement of about 3 Jy of unresolved source flux density at 1 m. Sources of this flux density should be only 5° or 6° apart on the sky so one can expect a suitable calibration source near any field under observation. For direct calibration at 4 m a source of about 25 Jy flux density is required to begin self-calibration. Almost every 3C source satisfies this requirement.

3.3. Isoplanatic Patch Size

Central to all discussions of calibration is the size of the ionospheric isoplanatic patch, i.e. the angular size of the region over which signals traversing the ionosphere maintain a constant phase relationship. This determines the angular distance that can be allowed between the unknown source being observed and the phase calibration source. The calibrator must be within the same isoplanatic patch as the source under observation. At shorter wavelengths the isoplanatic patch is much larger than the primary beam of the telescope and these considerations do not arise, while at 4 m the primary beam is some 13° and the isoplanatic patch is 3° or 4° across. If one is to image the whole primary beam, separate calibrations must be made for each isoplanatic patch.

If the isoplanatic patch is large we can be assured of having at least one suitable calibration source available, but this may not always be the case. At present, only preliminary measurements have been made of the patch size, indicating that it is normally a few degrees across. More data and experience are needed to better determine the patch size and how it may change with varying ionospheric conditions. We also need to produce a full list of suitable calibration sources for both wavelengths. Many more sources can be used than are listed in the present VLA Calibration Manual.

3.4. Diurnal and Solar Cycle Variations

It is often assumed that the best time to conduct 4 m observations would be at night during the years near the minimum of the solar cycle. This is not necessarily true, in fact, the opposite may be closer to the truth. Ionospheric densities are, indeed, much higher during daytime at solar maximum. The ionosphere can often absorb radio waves in the decametric range, but absorption at 4 m is almost always minimal. The important factor is the uniformity or smoothness of the ionosphere, not its peak electron density. The uniformity of the ionosphere governs the level of the phase fluctuations; the peak electron density controls the total phase shift, which is unimportant.

The ionosphere is most inhomogeneous late at night, around midnight or a few hours after midnight, when the final stages of recombination are taking place. This is often the poorest time for observing. Once recombination is complete and the electron densities are at their lowest levels, the ionosphere is usually stable again until dawn, when solar ultraviolet radiation causes a rapid build-up of ionization. Even though ionization levels are high during the daytime, the ionosphere is often quite uniform and excellent observing conditions ensue. However, poor ionospheric "seeing" can occur at any time; ionospheric weather is even less predictable than tropospheric weather.

With regard to the uniformity of the ionosphere, there appears to be little difference between solar maximum and solar minimum, but the available data are sparse and only anecdotal. Near solar maximum, large solar flares are much more common than near minimum. They can cause ionospheric disturbances which may result in such rapid phase fluctuations that calibration at 4 m would be impossible. Such severe disturbances usually last for only minutes or hours but, occasionally, they can last for days. Disturbances of this nature occur about once or twice per month near solar maximum and almost never during solar minimum. Therefore, with a varying amount of good luck, 4 m observations are quite feasible throughout the day and throughout the solar cycle.

4. Long Wavelength Observing

4.1. Angular Resolution

For extragalactic sources observations the rather poor, $25''$, resolution of the 4 m system represents its most severe limitation. It may be possible to partially alleviate this problem in the future by instrumenting the four VLBA antennas that are nearest to the VLA. This would provide an effective aperture of about 600 km and $\sim 1''$ resolution. Such a system would have rather low sensitivity, however, since only 10 baselines would be available instead of the 351 VLA baselines and observations would have to be limited to rather strong sources. Previous long wavelength VLBI work has indicated that it probably is not worthwhile to go to baselines $\gg 600$ km at 4 m; except for pulsars, all moderately strong sources appear to be highly resolved on these longer baselines either because of scattering effects or because of brightness temperature limitations in compact extragalactic sources.

The telescope sites proposed in the VLA Upgrade for the A+ system would be excellent locations for 4 m receivers. A moderately sensitive, high-angular-resolution system would then be formed.

4.2. Sensitivity

We are not yet certain about the ultimate sensitivity of the 4 m system and, in any event, it will certainly depend upon the direction of observation with respect to the Galactic Plane. The data are not fully reduced from the first observing run with the full 4 m system, but a map of the Coma Cluster area shown in Figure 3 has an RMS noise level of 25 mJy/beam after a 6.9 hour observation. In obtaining this map, no sophisticated interference flagging techniques were employed, nor were any wide-field deconvolution techniques used to reduce the noise caused by sidelobes of distant sources. Such techniques may bring the RMS noise down to ~10 mJy/beam.

4.3. Interference

Two forms of naturally occurring interference are well known. Lightning static will certainly be a major problem in the 4 m band during summertime observations. Also, dust particles pick up electrical charges under dry conditions and during high winds these charged particles striking an antenna cause large amounts of noise. Data can be seriously degraded by these phenomena.

Man-made interference is also a serious problem. However, the 4 m bandpass has been placed within the 73.8 MHz band exclusively allocated to radio astronomy. In our experience so far, this band is very quiet - we have not yet documented ANY external interference - so the band appears to be quieter than the 1 m or even the 20 cm bands. However, as is ironically the case with almost every radio observatory that attempts long wavelength work, on-site equipment at the observatory generates serious interference levels. Quite successful efforts have been made during the past few years to reduce these levels. The greatest problems were caused by radiation from the "B" racks in the vertex rooms of each telescope. These racks have now been placed in RFI shields that are very effective. Where possible, noisy digital equipment at the VLA site has been removed to the AOC. Problems remain with the 100 kHz oscillators in the bases of each telescope. These oscillators generate harmonics at 100 kHz intervals in the 4 m band and, unfortunately, they cannot be easily shielded. This problem is under discussion. At the present time it is alleviated by operating in the spectral line mode and flagging the frequency channels that are affected. The interference comes and goes, probably as oscillators drift in and out of synchronism, and does not appear on all baselines. The problem is most severe when interference is radiated by one telescope and received by a nearby one so that it then appears in the cross-correlated output from that baseline. Thus, the problem is greater in the compact configurations; it makes C and D-configuration observations difficult.

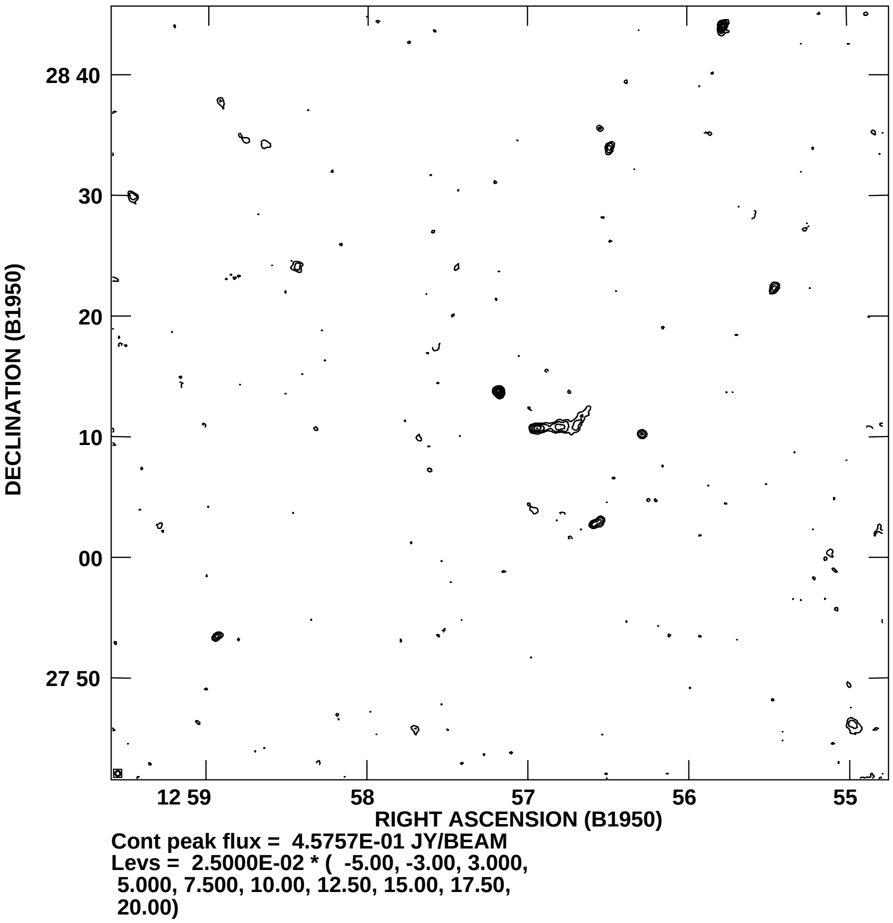


Figure 29–3. A 4 m (73.8 MHz) map of the Coma Cluster region incorporating only A-configuration baselines. The angular resolution is 25″. Contour levels range from -5 to 20 times the RMS noise of 25 mJy/beam. A head-tail radio galaxy is seen in the center of the map. The large-scale halo emission that is known to exist in the Cluster is resolved out with only A-configuration data.

4.4. Wide Field Mapping

For many programs it would be useful or even necessary to image the whole 13° primary beam at 4 m. This involves many problems, some of which are beyond the present capabilities of the software, especially for A and B-configuration work.

3-D Problem. This problem occurs for non-coplanar arrays and is normally handled by imaging the curved field using a large number of flat facets. In the 4 m case the number of required facets is larger than IMAGR can handle. This problem can be partly alleviated by relaxing various requirements and by limiting observations to high elevations. Cornwall's SDE task DRAGON has no limitation on the number of facets, but this task can be utilized on only a few specialized computers.

Isoplanatic Problems. Sources far from the field center produce sidelobes that must be removed from the central field for accurate imaging. In order to do this, these distant sources must be imaged. However, at 4 m these sources will be in different isoplanatic patches and a separate self-calibration solution will be necessary for each distant isoplanatic patch before the sidelobes produced by sources in that patch can be removed. Software has yet to be developed to accomplish this. Once again, this problem is significant only for the A and B configurations.

Bandwidth. To avoid bandwidth smearing of features far from the field center, the data will have to be averaged into 100 kHz bins, separately mapped, and the maps superimposed. The spectral-line software can do this automatically and it presents no particular problem.

5. GPS Data

Experiments have been conducted (Erickson et al. 1999) to evaluate the usefulness of ionospheric data obtained from the Global Positioning System (GPS) for the calibration of VLA data. Each GPS satellite radiates two L-band (1.4 GHz) signals that are coherent with each other. The delay between these signals provides a measurement of the Total Electron Content (TEC) along the path from the observer to the satellite. Simultaneously with VLA observations, GPS data were obtained from a receiver set up at the VLA site and/or one that is in regular operation at the Pietown, NM, VLBA site.

A model of the large-scale structure of the ionosphere over the VLA site was constructed from the GPS data. At any one time there are about 5 to 7 satellites at high enough elevations to be useful for ionospheric measurements. These provide TEC data at the points where the lines-of-sight to the satellites puncture the ionosphere. Since these puncture points are $\sim$500 km apart, we cannot model ionospheric structures $\sim$100 km in size; only those structures, such as large-scale wedges, with scale sizes $\geq$1000 km can be modeled. The model is quite adequate for the estimation of ionospheric Faraday rotation since the rotation depends only on the TEC (and the geomagnetic field). However, ionospheric phase fluctuations depend upon the differences in TEC along the ray paths from the source to the interferometer elements. These differences depend on both the large-scale structures and the smaller scale ones. Phase correction via the GPS data reduces the phase fluctuations by $\sim$ 50%. This is probably a useful procedure which will allow self-calibration to proceed more efficiently but it is only an aid to self-calibration, not a solution for the phase fluctuation problem.

Eventually we hope to routinely incorporate GPS data and ionospheric corrections based upon them into the VLA on-line observing system. Faraday rotation correction and partial correction of the phase fluctuations would then be made automatically.

References

Blundell, K. M., Rawlings, S., Eales, S. A., Taylor, G. B. & Bradley, A. D. 1998, *MNRAS*, 295, 265–279.

Cane, H. V. 1979, *MNRAS*, 189, 465–478.

Cordes, J. M. 1990 in *Low Frequency Astrophysics from Space* eds. N. E. Kassim & K. W. Weiler (Berlin: Springer-Verlag), 165–174.

De Breuck, C., Brotherton, M. S., Tran, H. D., van Breugel, W., & Rottgering, H. J. A. 1998, *AJ*, 116, 13-19.

Erickson, W. C. 1964, *ApJ*, 139, 1290–1311.

Erickson, W. C., Perley, R. A., Flatters, C., & Kassim, N. E., 1999, *A&A*, (in preparation)

Fey, A. L., Spangler, S. R. & Cordes, J. M. 1991, *ApJ*, 372, 132–160.

Garriott, O. K., Smith III, F. L., & Yuen, P. C. 1965, *Planet. Space Sci.*, 13, 829–838.

Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott P. F., & Collins R. A. 1968, *Nature* 217, 709–713.

Hewish, A., Scott, P. F., & Wills, D. 1964, *Nature* 203, 1214–1217.

Kassim, N. E., Perley, R. A., Erickson, W. C., & Dwarakanath, K.A. 1993, *AJ*, 106, 2218–2228.

Lawrence, R. S., Little, C. G, & Chivers, H. J. A. 1964, *Proc. IEEE*, 52, 4–27.

Landecker, T. L. & R. Wielebinski 1970, *Aust. J. Phys. Suppl.*, 16, 1–30.

Scott, S. L., Coles, W. A., & Bourgois, G. 1983, *A&A*, 123, 207–215.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

30. Pulsar Observing at the VLA

T. H. Hankins
New Mexico Institute for Mining and Technology, Socorro, NM 87801, U.S.A.

Abstract. Pulsar observing at the VLA can take advantage of some of the unique attributes of the array. Several areas of scientific endeavor are outlined, the available facilities for pulsar observing are described and some suggestions for observing preparation are listed.

1. Introduction

The Very Large Array has several attributes well-suited for several types of pulsar observing. It has the largest collecting area of any telescope outside of the Arecibo Observatory's declination range, the array can be operated simultaneously on several pairs of frequencies, the synthesized beam can be used to isolate pulsars in bright background regions such as the Crab Nebula, and the array can be operated over a wide frequency range; at the VLA pulsars have been detected from 74 MHz to 22.5 GHz with declinations as low as $-47°$.

In this lecture I describe some of the types of observations that have been conducted at the VLA, some of the limitations on observing, the equipment available for data acquisition, and the general procedures for pulsar observing, some of which depart from standard synthesis imaging operations.

2. Scientific Objectives of Pulsar Observing at the VLA

A wide variety of pulsar investigations has been carried out at the VLA. Some of these are listed below with the expectation that these ideas may spark new ones not yet tried. One way to categorize VLA observations is by the type of data product desired. I divide these into time series, spectral, and image products. The first two categories produce data which are similar to what may be obtained from a single dish, yet they can take advantages of particular characteristics of the VLA. Then one uses the VLA for its large collecting area and sky coverage (70% of the celestial sphere). Other projects require the high angular resolution and synthesis imaging capabilities of the VLA.

2.1. Time Series

Pulsar Timing. During the time that the Arecibo Observatory has been out of commission for its upgrade, the VLA has been used to measure the times-of-arrival of several pulsars of special interest which lie in the Arecibo declination range. Other pulsars outside of the Arecibo declination range have been timed on a regular basis for many years. From precision pulsar timing one can obtain the period *vs.* period derivative plot for pulsars, often called the pulsar HR diagram. Binary pulsars provide the best laboratory for testing the theory of General Relativity; the "Holy Grail" for those who search for new pulsars is to find one in orbit around a black hole. The resulting mass determination would provide unambiguous evidence for the existence of a black hole.

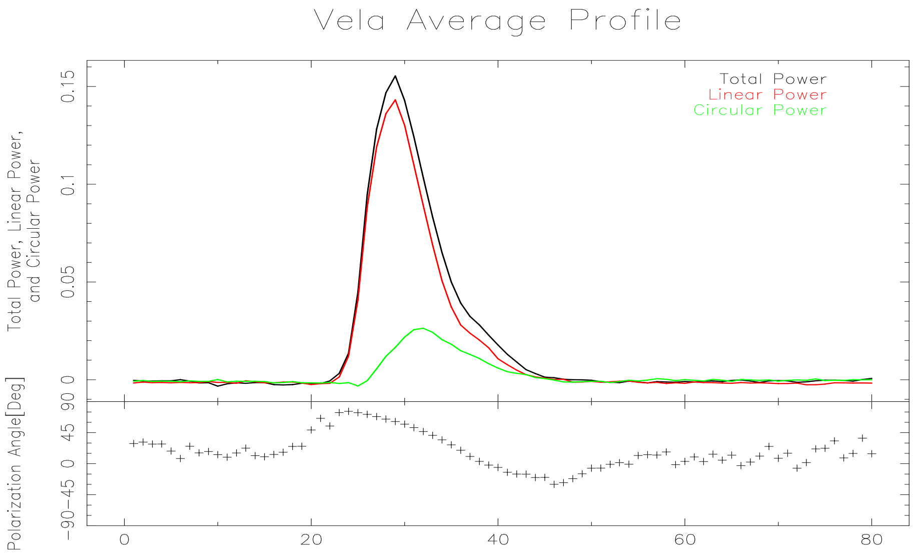


Figure 30–1. Polarization of the Vela pulsar recorded at the VLA at 1420 MHz. The linear polarization fraction is nearly 100% at the center of the profile. The position angle rotates smoothly across the pulse window. This has been interpreted as mapping the projection on the celestial sphere of the diverging magnetic field lines from the magnetic polar cap.

Directed Searches. Searching for radio intensity periodicities with the synthesized VLA beam is practical only for known objects. Both X-ray pulsars and steep spectrum radio sources have been candidates for pulsar searches.

Polarization of Individual Pulse Sequences and Average Profiles. Polarimetry is probably the best diagnostic we have for pulsar magnetospheric geometry. An example is shown in Figure 30–1. Polarization and high time resolution intensity studies are essential for understanding the pulsar emission mechanism.

The Crab Nebula. When the VLA is operated in its phased-array mode, the synthesized pencil beam is so small that when directed at a point-source pulsar embedded in a bright supernova remnant (SNR), the bright background is spatially resolved. This technique has been used extensively to study the Crab Nebula pulsar and its giant pulses. When using a large single dish telescope with a beamwidth of more than a few arc minutes, the Crab Nebula typically increases the system temperature by an order of magnitude. For example, when pointing the Arecibo telescope at the Crab pulsar the system temperature at 430 MHz increases to more than 20,000 K, thus increasing the detected signal fluctuations by a factor of 200. At the VLA the Crab Nebula increases the L or P-band system temperature of each antenna by about 100 K, but since the Nebula is spatially resolved, it does not contribute to the correlated point source flux.

Hence the obtainable S/N on the Crab pulsar is vastly better than that which can be obtained at a single dish.

By resolving out the nebula background we have found anomalous components of the Crab pulsar at 1.4, 4.8 and 8.4 GHz and studied their polarization, and we have found unresolved nanostructure at 4.8 and 8.4 GHz with a duration of 10 ns. If one makes the conventional assumption that the size of an emitting region cannot be any larger than the time it takes for light to travel across it, we have found radiating entities whose dimensions are $\approx 3\,\mathrm{m}$, with brightness temperatures exceeding 10^{31} K (Hankins 1996, Moffett 1997).

2.2. Imaging

Pulsars are unresolved point sources unless their images are broadened or distorted by interstellar propagation. Imaging experiments are therefore conducted either to determine a pulsar's position precisely, to image the neighborhood of a pulsar, or to measure the interstellar scattering effects.

Astrometry. Several groups are interested in precision astrometry of pulsars for proper motion determinations (*e.g.*, Fomalont et al. 1984, 1992). Although several proper motion studies have been made using the VLA alone, the next step is clearly to use the VLBA for parallax and proper motions of more distant pulsars. To optimize the signal-to-noise ratio (S/N), the VLBA correlator must be gated synchronously with the pulsar to eliminate the noise contribution to the correlations during the typically 90% of the time during which a pulsar is "off". This requires an up-to-date timing model so that the pulse phase can be accurately predicted by the VLBA correlator. Since the most sensitive VLBA observations include the phased VLA as one of its stations, simultaneous timing measurements of pulsars can provide unambiguous timing models referred to the VLA master clock.

Precise position determination of pulsars using the VLA has been used to improve the timing model solutions for several pulsars soon after their discovery. In one case the VLA was gated synchronously with the pulsar at a 50% duty cycle, then the "on" and "off" images were compared to see which point source in the imaged field was present in only one of the images. In another case the improved VLA position of the pulsar enabled sufficient refinement of the timing model that the residuals showed the existence of several planetary mass objects in orbit around the pulsar (Wolszczan & Frail 1992).

Unpulsed Emission from Pulsars. With a single-dish telescope it is impossible to determine if the pulsar flux goes to zero between pulses. However, by synchronously gating the correlator and making separate synthesis images when the pulsar is "on" and when it is "off", an accurate intensity baseline can be determined for the "off" gated image. For one pulsar it has been shown that the suspected continuous emission may be due to an unresolved point source 13.5 arcseconds away from the bright pulsar B2028+16 (Hankins et al. 1993). In a study of the Vela pulsar Bietenholz, Frail, & Hankins (1991) used the "on/off" gating to study weak continuum emission from the nearby region of Vela by gating the correlator "off" during Vela's strong pulse.

Solar Wind Studies. The VLA can synthesize a pencil beam; when pointed near the Sun the effect on the beam due to the Sun being spatially resolved is reduced, relative to the effect on a single-dish telescope. Then the scintillation, dispersion measure changes and Faraday rotation of signals from a pulsar passing near the Sun should be measurable. Since the data can simultaneously be imaged, some statistical estimates of coronal structure near the Sun can be made. The coronal electron density can be estimated from the additional dispersion delay, and the longitudinal component of the coronal magnetic field can be estimated from changes in Faraday rotation of the pulsar's linear polarization.

2.3. Spectral Analysis

Galactic HI Emission and Absorption. Since pulsars are alternately "on" and "off", both the absorption and emission spectra are "simultaneously" available allowing resolution of the HI distance ambiguity (Frail et al. 1991) inside the solar galactic circle. For these measurements the VLA correlator is gated but in a mode where the "on-pulse" and "off-pulse" spectra are separately accumulated.

Scintillation and Dynamic Spectra. The time and spectral resolution of the VLA correlator allows the measurement of pulsar dynamic spectra. From the scintillation patterns produced, estimates of the structure and relative velocities of the interstellar medium can be made.

3. Observing Issues

For pulsar observations there are a number of limitations and pitfalls of which one must be aware. Some of these are described briefly here and others may be found in the *A Guide to Pulsar Observing at the VLA*, (Moffett & Hankins 1998).

Proper Motion. The proper motion of some nearby pulsars is sufficient that one must use the current position of date precessed to the observe file epoch (B1950.0 or J2000.0). The bright pulsar B1133+16, for example, has a proper motion of 357 mas yr^{-1} in declination, so it moves a 21 cm A-array beamwidth in a few years. In the phased-array mode, if the position is not corrected, a bright pulsar may be seen only very weakly and sporadically as it passes through the dirty-beam sidelobes; a weak pulsar will not be seen at all.

Waveguide Switch Cycle. The IF signals from each antenna are frequency-multiplexed onto a buried waveguide that transports the signals to the VLA Control Building. The same waveguide is used to transport control, local-oscillator reference, and engineering information to the antennas. To accomplish this the antennas transmit data to the Control Building for about 50 ms (This interval is called "data valid."), then the direction is reversed and information is sent to the antennas for about 1.5 ms. During this 1.5-ms interval (called "data invalid", which occurs at the waveguide cycle repetition frequency of 19.6 Hz), there is no sky signal available. Without compensation this data interruption would preclude any pulsar observations which rely on the conventional square-law detection and modestly rapid sampling; during the 1.5-ms "data invalid" time, the detected output falls towards zero, producing a strong transient at the start and end of the interval. The VLA square-law detectors and polarization cross multipliers alleviate most of this problem by incorporating "track-and-hold" circuits which latch the detected signal level at the end of the "data valid" interval, and hold this value until sky data is available again. These circuits are controlled by a logic signal called DATA_VALID which is asserted for about 50 ms during the fraction of the waveguide switch cycle that data from the receivers is available.

There are two consequences of the data "dropout" during the data invalid interval. One is that spectral pulsar searches are likely to have some contamination at 19.6 Hz and its harmonics. We generally minimize this by sampling continuously at some integer multiple of 19.6 Hz (and 60 Hz), so that all spectral components of the waveguide switch cycle (and harmonics of 60 Hz) fall exactly into Fourier transform spectral bins. The other effect is that computed pulse profiles are slightly broadened and shifted toward later times when track-and-hold is asserted during a pulsar pulse. This biases the true arrival time by a small (but consistent) amount. The logic state of the DATA_VALID signal is recorded with the data, and thus unbiased average profiles can be computed by discarding samples taken when DATA_VALID is deasserted and counting the valid samples in each phase bin. All of the VLA real-time signal averagers perform this correction.

For normal operation a known amount of thermal noise is injected into the receivers every other waveguide cycle. Since this would also disrupt pulsar observing, the noise tubes are turned off for pulsar observations by requesting in advance that the VLA operator remove the appropriate flags in the Subreflector and Front End Parameters file (better known as the SYS*b*ROT file, where *b* is the band designator P, C, L, X, *etc.*). These flags are in columns 64 and 65. It is also normal practice to disable the correlator self-tests (which normally occur during the "data invalid" interval) by requesting that the VLA operator place a `T` in column 15 of the third card of the ARRAY file when gating the correlator.

Limits on Correlator Gating. The VLA correlator operates on a 92.8 μs integration cycle which is phase-locked to the station standard frequency. To avoid incomplete integration cycles, the pulsar correlator gate is synchronized to the correlator integration cycles by the "Correlator Clock Box", located in the correlator room behind the correlators. The Clock Box maintains proper synchronization by delaying the start of the gating signal until the computation of a fresh set of lead/lag products has begun. But since the correlator uses recirculation to expand its spectral resolution, the temporal quantization of integration cycles depends upon the IF bandwidth. With the full 50 MHz bandwidth the recirculation factor is unity, so the gate granularity is the fundamental 92.8 μs. For narrower bandwidths, however, the granularity increases so that for 1.56 MHz bandwidth the gate may be delayed by up to 2969.6 μs. The result is a jitter between the relative phase of the pulsar correlator gate and the actual correlator computation window. For narrow pulsars and those with short periods, this "jitter" may be unacceptable. The control of the Correlator Clock

Box is strictly manual. The recirculation code must be manually set by a rotary switch on the Correlator Clock Box to match the IF bandwidth.

Dynamic Range Limits of Mark III VLBI Video Converters. The Mark III VLBI Video Converters were originally designed for VLBI work where the signals are split into separate frequency channels, then sampled with two or three levels of quantization. Therefore linearity and dynamic range were not a high priority in its design. The consequence is that the input levels have to be fairly carefully set to avoid underflow noise or overflow saturation. The problem is compounded because each video converter filter has a slightly different gain, and there is no easy way to adjust them independently. Both the continuous sampling and burst sampling data acquisition programs continuously monitor and display the ADC voltages for all active channels.

Polarimetry Calibration. Calibration of the pulsar polarimeter is complex. The Stokes' parameters of a linearly polarized calibrator are measured at a range of parallactic angles, then sinusoids are fitted to the Stokes parameters as a function of parallactic angle to obtain the ellipticity and coupling coefficients. The detector and cross-multiplier gains are obtained either by measuring a linearly polarized source or by synthesizing one with quadrature hybrids. On strong pulsars one can do a polarimetric "self-calibration" by choosing a specific pulse longitude where the linear polarization is strong and determining both the ellipticity and the polarimeter gains from the fluxes at this pulse longitude. (McKinnon 1992, Moffett 1997)

Data Analysis Software. There is no pulsar-specific data analysis software supported by NRAO.

4. NRAO Supported Equipment

NRAO supports the capability to form the Stokes' parameters from two IFs and record the detected, smoothed quantities at sample intervals down to about 160 μs per channel. There is also an implementation of the Dartmouth/Princeton Mark 3 Timing System available, and provision for straightforward connection of user provided equipment. Many of the signals used for pulsar work are brought to a BNC patch panel near the center of the pulsar equipment racks for easy monitoring or experiment modification. A block diagram showing some of the interconnections between elements of the pulsar equipment at the VLA is given in Figure 30–2. The functional blocks are described in the following sections.

Analog Sums. The Analog Sums are the equivalent of the IF outputs of a single-dish telescope. There are four Analog Sum signals, corresponding to the four VLA IFs, A, B, C, and D. The Analog Sums are formed by converting the 3-level sampled outputs of the interferometer delay lines to analog voltages and then summing these voltages first for each array arm, then for the whole array. They contain frequencies from about 25 kHz to the specified array IF bandwidth. If, for example, the VLA is observing with a 50 MHz bandwidth

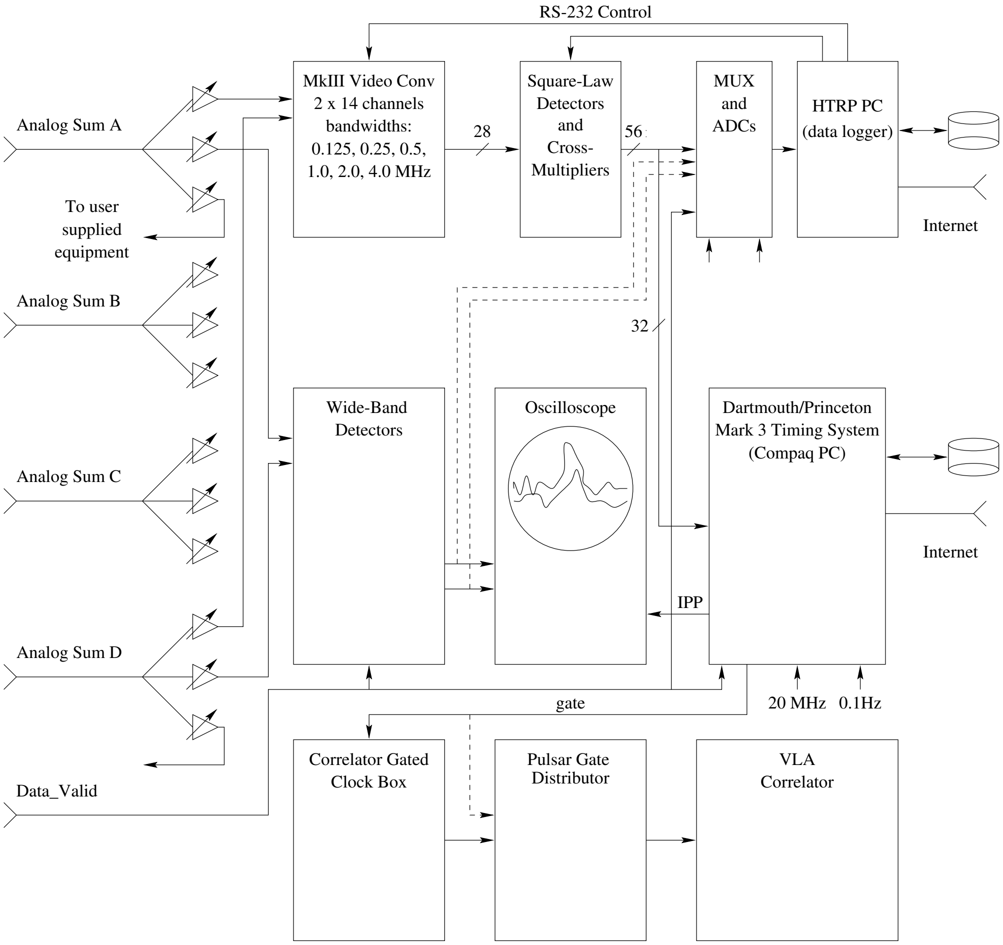


Figure 30–2. Block diagram of the essential elements of the pulsar data acquisition system.

centered at 1414.9 MHz, then the sky center frequency of 1414.9 MHz appears in the Analog Sum at 25.0 MHz and the upper bandedge of 1439.9 MHz appears at 50 MHz. The frequency order is inverted for those bands, 8 GHz and 15 GHz, for which the local oscillator is on the high side.

The Analog Sums are generated in the Correlator Room and are sent to a distribution panel where they are separately buffered for VLBA and pulsar use. They are further split and isolated by the Pulsar IF Distribution System, which allows each IF to be split into three independent paths, usually to the Wide-Band Detector, to the Mark III Video Converters, and to a user equipment port.

IFs A and D are usually operated as an orthogonally circularly polarized pair. IFs B and C can also be used as a polarization pair, but only one pair can be "phased up" at a time. Specifying "autophasing" (or `VA`) mode in the observe file phases up IFs A and D.

Wideband Detectors. A pair of wideband detectors is available, which can accept the full Analog Sum bandwidth. It contains a 0–4 db input attenuators, adjustable time constants, ($0.02\,\mathrm{ms} \leq \tau_{\mathrm{RC}} \leq 10\,\mathrm{ms}$), and output DC offsets ($0.0\,\mathrm{V} \leq v_{\mathrm{offset}} \leq +10.0\,\mathrm{V}$), plus input and output level monitors. The square-law detector outputs are "frozen" or held during each waveguide switch data interruption when DATA_VALID is deasserted. This removes the strong transient which would otherwise occur. It also contains circuitry to compute a 3-second running mean of the *rms* detected noise and comparator circuitry to indicate when the detected signal exceeds n times the *rms* noise, ($1 \leq n \leq 8$). n is adjustable in integer steps. This feature is convenient both for triggering on large pulsar pulses and for flagging strong impulsive interference. The outputs of the Wideband Detectors are commonly monitored by an oscilloscope triggered synchronously with the pulsar for data quality monitoring.

Mark III VLBI Video Converters. The Mark III Video Converters (VC) are used as a dual 14-channel filter bank to aid in dispersion compensation. The Analog Sums are mixed with a 300 MHz local oscillator and then sent to the Video Converter IF Distribution panel where the signals are sent to each VC. Each channel has a full set of filters, allowing channel bandwidths to be set to 0.125, 0.25, 0.5, 1, 2, or 4 MHz. Each *pair* of VCs has its own local oscillator, so pairs of channels can be placed anywhere in the array passband. A pair consists of an upper side band (USB) and lower side band (LSB). The effective center frequencies of the USB and LSB filter bank channels, $f_{\mathrm{Sky}_{\mathrm{USB}}}$ and $f_{\mathrm{Sky}_{\mathrm{LSB}}}$ are

$$\begin{aligned} f_{\mathrm{Sky}_{\mathrm{USB}}} &= f_{\mathrm{VLA\ center}} + (f_{\mathrm{VCLO}} - 300.0\,\mathrm{MHz}) - \frac{1}{2}(\Delta f_{\mathrm{VLA}} - \Delta f_{\mathrm{VC}}), \\ f_{\mathrm{Sky}_{\mathrm{LSB}}} &= f_{\mathrm{VLA\ center}} + (f_{\mathrm{VCLO}} - 300.0\,\mathrm{MHz}) - \frac{1}{2}(\Delta f_{\mathrm{VLA}} + \Delta f_{\mathrm{VC}}), \end{aligned}$$

where $f_{\mathrm{VLA\ center}}$ is the sky center frequency to which the VLA is tuned, f_{VCLO} is the frequency to which the VC LO is set, Δf_{VLA} is the VLA IF bandwidth, and Δf_{VC} is the bandwidth of the VC channel. Both f_{VCLO} and Δf_{VC} are set from the VC front panel.

Although the VC LOs are phase-locked to the station standard maser reference frequency, they are not phase coherent with one another. Thus if the frequencies are changed or power to the VCs is interrupted, the relative phases of the output signals is changed. The consequences of this are most serious in polarization calibration; no changes to the VCs can be made between observations of polarization calibrators and a pulsar.

Square-law Detectors and Cross-multipliers. The VC outputs are sent to a set of square-law detectors and cross-multipliers. Sample-and-hold circuits are provided at the outputs so that the waveguide switch transient is avoided and so that all channels can be sampled at the same epoch. The detector time constants can be set to 25, 50, 100, 200, 500, 1000, 2000, or 5000 μs, and the output can be offset up to ±5 V through a RS-232 connection to the data-logging PC.

If the VLA IF signals are nominally right (R) and left (L) circular polarization, the detector outputs are LL, RR, LR, and $LR(\pi/2)$, where the latter indicates that the right circular polarization is shifted 90° in phase prior to multiplication. The cross-products, LR and $LR(\pi/2)$ are also known as *RLCOS* and *RLSIN*, respectively. Thus the outputs of the polarimeter can be used to form all of the measured Stokes parameters for the independent channels:

$$\begin{aligned} I &= (LL + RR)/2, \\ V &= (LL - RR)/2, \\ Q &= RLCOS, \\ U &= RLSIN. \end{aligned}$$

HTRP Data Logger. For historical reasons the pulsar data logger has become known as the HTRP (High Time Resolution Processor). In fact its time resolution is modest; currently (1998) it can continuously sample 64 input channels at intervals of about 160 μs, for an aggregate data rate to disk of about 200 samples/s. In “burst mode” it can sample several times faster, so long as the aggregate data rate maximum is not exceeded.

The HTRP consists of a PC containing four 16-channel analog-to-digital converter cards and about 10 GB disk space. The ADCs are triggered either by a computer-controlled synthesizer referenced to the station frequency standard, or by an external TTL-level source, as is used in the “burst mode”, where the sample clock and burst gate are typically obtained from the Dartmouth/Princeton Timing System.

The HTRP data acquisition programs can be run under Windows NT, but for highest performance, MSDOS is preferred. Utilities for setting the detector time constants and DC bias, for monitoring detector levels, and for data archiving are available.

5. Non-supported Equipment

Dartmouth/Princeton Timing System. The Dartmouth/Princeton Mark 3 Timing System (DPTS) consists of three custom ISA cards that mount in a PC, and an extensive MSDOS software suite (Stinebring et al. 1992). It synchronously samples up to 32 detector channels with 6-bit resolution at intervals down to 12.8 μs per channel and computes average profiles for each channel and a dedispersed average profile, time tagged to 0.1 μs resolution. It has been used extensively for pulsar timing by the Princeton Pulsar Group at Arecibo, Green Bank and the VLA.

The DPTS has the provision for setting two independent, pulsar-synchronous windows or gates at arbitrary pulse phases. These windows are available as TTL logic signals that can be used for gated sampling by the HTRP and for gating the VLA correlator synchronously with the pulsar.

Dartmouth Signal Averager. The Dartmouth Signal Averager can be used to display a real-time pulsar average profile of a single detected channel and generate a pulsar synchronous gate. It can be used as a backup gate generator for non-binary pulsars in case of failure of the DPTS.

Future Instrumentation. The pulsar data acquisition systems will continue to improve, particularly in time resolution. At this writing a new 32-channel single-pulse and averaging system in a VME form factor is under development. It will allow more convenient remote operation, more versatile data displays and extensibility to incorporate real-time digital signal processors. A system to record the full bandwidth of the Analog Sum voltage before detection is also under development by New Mexico Tech. This system will allow much higher time resolution and more flexible interference excision at the cost of extensive processing for dispersion removal.

6. General Procedures

Preparation of Observe File. In the program `observe` with which you create an observe file, in the startup screen under `Special Instructions` enter something like, "Observer will be present at the VLA to operate pulsar equipment. Please create subreflector rotation file THHLROT and turn off noise tubes by placing a 'T' in columns 64 and 65 of the file." This sentence then appears as a comment at the beginning of the observe file. The array operator will make the special ROT file in advance of your observing time. For the name of the specially requested ROT file I usually use my initials, followed by the band letter, followed by 'ROT', *e.g.*, `THHLROT` instead of the standard default file `SYSLROT`. Type in this name after the item `FE/Subref` on the appropriate source card screen the `observe` program.

Observe files for phased-array pulsar observations require first pointing the array at a known point-source calibrator near (preferably $< 15°$ from) the pulsar in VA mode to determine the antenna peculiar phases. Then the array can be moved to a flux or polarization calibrator, blank sky or a pulsar using the VX mode (called "apply last phase" in the `observe` program). In the case of a pulsar, correction of position for proper motion should be considered. More information about phasing the array can be found in Wrobel & Taylor (1997).

The array usually remains phase-stable for up to an hour or more, except at low frequencies when the ionosphere can change more rapidly, especially at dawn. If you are in doubt as to how well the array has remained "phased up", you can point again at the calibrator in VX mode and display the current phases on the array operator's console.

Preparation of `polyco.dat` File for Synchronous Gating or Windowing. The DPTS requires a timing model for the pulsar of interest. The program `TEMPO`, which is supported by the Princeton Pulsar Group and is publicly available (*http://pulsar.princeton.edu/tempo/index.html*), can be used to generate a file named `polyco.dat` of polynomials in pulse phase as a function of observing time. The DTPS evaluates this polynomial to produce the current Doppler-shifted pulse frequency for synchronous sampling.

Determination of Pulse Phase Using the Dartmouth/Princeton Timing System. The DPTS displays the average waveform of the pulsar period. One of the command line options allows setting both the delay from the start of the pulse period to the start of a synchronized logic window and the duration of the window. At the next integration cycle these window boundaries are displayed and the synchronized logic window is turned on. This signal can be sent to the VLA correlator to gate it synchronously with the pulsar. There are several switches that must be set in the correlator room to enable the correlator gating. These switches are described in Moffett & Hankins (1997).

Operation of HTRP, Options for Monitoring Data and Data Archiving. The HTRP computer operation is straightforward. There is a program for initialization of the detector time constant and bias, and several for data acquisition in different sampling modes. Data are normally recorded on a set of disk drives and then archived to tape in `tar` format after an observing session. For short runs it is often feasible to transmit the data *via* `ftp` to a remote archive machine.

During operations the detector output voltages are displayed on the HTRP computer screen for monitoring purposes. The display sparsely samples the input data, updating the screen about once per second, but this serves to indicate gain and bias levels and slowly changing interference.

Requirement to be on Site. The time to phase the VLA on a point source calibrator is variable, as is the scintillation intensity time scale for pulsars, especially at high frequencies. Hence pulsar observing tends to be quite interactive with frequent requests to the Array Operator to step from one source to the next in the observe file. Furthermore, the pulsar equipment has not been designed for remote operation. So the pulsar observer should plan on being at the VLA site for his observations, and should be prepared for a busy, interactive session.

7. Documentation

As the pulsar data acquisition systems at the VLA are under more or less constant development, the documentation rapidly falls out of date. However, as part of his Ph.D. Thesis, David Moffett wrote the most recent handbook, which I have recently updated. It should soon be available on the NRAO web page as *A Guide to Pulsar Observations at the VLA*, by D.A. Moffett and T.H. Hankins.

References

Bietenholz, M. F., Frail, D. A., & Hankins, T. H. 1991, *ApJ*, 376, L41–L44, (1991).

Fomalont, E. B., Goss, W. M., Lyne, A. G., & Manchester, R. N. 1984, *MNRAS*, 210, 113.

Fomalont, E. B., Goss, W. M., Lyne, A. G., Manchester, R. N., & Justtanont, K. 1992, *MNRAS*, 258, 497.

Frail, D. A., Cordes, J. M., Hankins, T. H., & Weisberg, J. M. 1991, *ApJ*, 382, 168.

Hankins, T. H. 1996, in *Pulsars: Problems & Progress*, ASP Conference Series, S. Johnston, M. A. Walker, & M. Bailes, eds., 105, 197.

Hankins, T. H., Moffett, D. A., Novikov, A. & Popov, M. 1993, *ApJ*, 417, 735.

McKinnon, M. M. 1992, *Ph. D. Thesis* New Mexico Tech, Socorro, NM.

McKinnon, M. M. & Hankins, T. H. 1993, *A&A*, 269, 325.

Moffett, D. A. 1997, *Ph. D. Thesis* New Mexico Tech, Socorro, NM.

Moffett, D. A., & Hankins, T. H. 1998, *A Guide to Pulsar Observations at the VLA*, NRAO HTRP Memo Series.

Stinebring, D. R., Kaspi, V. M., Nice, D. J., Ryba, M. F., Taylor, J. H., Thorsett, S. E., & Hankins, T. H. 1992, *Reviews of Scientific Instruments*, 63, 3551.

Wolszczan, A. & Frail, D. A. 1992, *Nature*, 355, 145.

Wrobel, J. M. & Taylor, G. B. 1997, `http://www.nrao.edu/vla/vlbivla/vlbivla/vlbivla.html`.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

31. Solar System Objects

B. J. Butler and T. S. Bastian
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. Long wavelength ($\lambda \gtrsim 350\mu$m) interferometric observations of solar system objects can yield important information on the nature of these bodies, including information on orbits, spins, surfaces, atmospheres, magnetic fields, rings, and plasma processes. This lecture will describe some of the peculiarities involved with observing of solar system bodies with synthesis arrays. These include planning, scheduling, calibration, and imaging.

1. Introduction

Interferometric techniques have been used to observe the Sun and planets for as long as interferometers have been operating. In fact, the first radio interferometric observations were those of McCready *et al.* (1947), who used a single antenna on a cliff above the sea near Sydney, Australia, to observe the Sun. By observing the Sun as it rose in the east, a single antenna could observe both the direct radio waves and those reflected from surface of the ocean. The direct and reflected radio waves interfered with each other, producing the characteristic fringe pattern of an interferometer with a baseline roughly twice the height of the cliff above the sea. This "sea interferometer" first resolved discrete features associated with radio emission from the Sun.

Fifty years later, radio interferometric observations of the Sun, planets, and other solar system objects continue to make significant contributions to our understanding of these bodies. High resolution radio observations offer the means of obtaining information on the physical nature of solar system objects that is either unavailable through other observational techniques, or is highly complementary to the information obtained by other means. For example, radio waves can be used as an effective probe of the outer layers of the atmosphere of the Sun. Similarly, radio waves are unique in their ability to probe into the subsurfaces (regoliths) of the solid body planets, and into the deep atmospheres of the planets with significant atmospheres.

In broad terms, observations of solar system objects with a synthesis radio telescope are similar to those of any other celestial object: we use an ensemble of interferometers to measure the Fourier transform of the angular distribution of radio brightness from the Sun, the Moon, a planet (or its satellite), an asteroid, or a comet. However, unlike most celestial sources, solar system objects display significant changes in angular position during the course of an observation. Many solar system objects are strong and/or spatially complex at radio wavelengths. Furthermore, the radio emission from some objects may change significantly on short time scales (e.g., the Sun produces shortlived radio bursts; Jupiter rotates relatively rapidly). These factors have an impact on all levels of activity associated with observing: planning, observing, and post-processing.

The structure of this lecture is therefore as follows. We begin with some general considerations that apply to essentially all solar system objects. We then consider two special topics: interferometric radar observations of solar system objects and solar observations. Each of these major sections will include an example illustrating the quality of images of solar system objects which can be obtained.

2. General considerations

2.1. Source motion

With the exception of solar observations (see section 4.), the biggest difference between observations of solar system objects and those of other celestial sources is that source motion must be accounted for. Unlike these other sources, which are essentially fixed on the celestial sphere, solar system objects display both real and apparent motions. These motions are a combination of three types: 1) *Horizontal parallax*: A nearby source viewed from Earth appears to move against the background stars as the observer, fixed to the surface of the Earth, rotates from west to east. The magnitude of the parallax is $HP \sim 8.79''/D$, where D is the geocentric distance of the object in astronomical units (1 AU $\sim 1.496 \times 10^{11}$ m is the mean distance from the Sun to the Earth). 2) *Orbital motions*: Far larger than the horizontal parallax are the apparent motions resulting from the relative orbital motions of the Earth and a solar system object. Typical angular rates of motion for solar system objects are: 13 deg/day for the Moon; 1 deg/day for the Sun; 35 arcmin/day for Mars; 8 arcmin/day for Jupiter; and 2 arcmin/day for Uranus. Nearby objects will, of course, have higher rates of motion. Comets and asteroids which pass near the Earth can have rates as high as 1 arcsec/sec, or 24 deg/day! 3) *Intrinsic motion*: All solar system objects have an intrinsic component to their motion: rotation. This particular motion must be accounted for if the body being observed is larger than the primary beam of the antennas, and a particular location on the surface of the body is to be tracked. In practice, this is only done for the Sun, where the (synodic) period is 27.3 days.

In addition to horizontal parallax, orbital motion, and source rotation, solar system objects often display intrinsic variability in their radio emission. For example, the cloud layers on a planet slowly change in time, or the radio emission from the Sun can change quite rapidly.

2.2. Scheduling and planning

Several factors complicate the planning of observations of solar system objects. Many of these can be attributed to source motion and/or variability of the source. However, the motions of the Sun, the Moon, and the planets, and many asteroids, are well known. In this respect, favorable observing times can be anticipated well in advance of a particular experiment.

Most general purpose interferometric arrays go through fixed antenna configurations in order to sample a greater range and number of spatial frequencies. This works well for most sidereal sources because they do not vary on short timescales. Solar system objects do vary on a timescale that is short compared with typical array reconfiguration cycles. Hence, unlike observers of most sidereal sources, observers of solar system objects cannot exploit multiple array configurations to better sample the (u, v) plane. The viewing geometry – i.e., the subearth latitude and longitude, and the distance of the source – will be differ-

Table 31–1. Apparent sizes of large solar system bodies

body	min diameter (″)	max diameter (″)
Mercury	4.6	12.3
Venus	9.7	63.4
Mars	3.5	25.1
Jupiter	30.6	49.9
Saturn	15.0	20.6
Comet Hyakutake	0.03	4700
Comet Hale-Bopp	0.05	360

ent from configuration to configuration, and the body itself has likely changed. This means that before the observations are proposed, the goals of the observation must be well determined. The expected visibility function for a disk (i.e., planetary) source is derived in Appendix A, where it is shown that the spatial frequency parameter of interest (β) is the product of the apparent radius of the body and the projected spacing between antennas in wavelengths. If information on the large scale structure of the body is the goal, then observations must be undertaken when small spatial frequencies are adequately sampled ($\beta \lesssim 0.5$ or so). Table 31–1 shows the sizes of the larger bodies in the solar system, excluding the Sun and Moon (which are both about 30 arcminutes in diameter). Also shown in Table 1 are two recent comets (for the comets the sizes are at perigee - min size is the nucleus, max size is the approximate extent of the OH coma). It is easy to see that small antenna separations are necessary to obtain good large scale structure information on these bodies. If the small scale structure is the goal, then observations must be undertaken when the antenna configuration allows higher spatial frequencies to be sampled.

There is often a mismatch between the time an optimum observing geometry is available and the time a favorable array configuration is available to meet the scientific goals of a proposed observation. For example, if the highest possible resolution on Mars is the desired goal for an observation with the VLA, then observations in the A configuration at a time near opposition (when Mars is closest to Earth) are best. However, oppositions of Mars are spaced by about 2 years, meaning that not all oppositions can be observed with the VLA in the A configuration. If it is possible to move the antennas on short time scales (days), then this is not as much of a problem. This is the case, for example, at the Owens Valley Radio Observatory millimeter array, but such rapid antenna reconfigurations place a strain on observatory operations.

If the object is a rapid rotator (e.g., Mars, Jupiter or Saturn), then the data must either be averaged together into a longitudinally smeared result, or, if sufficient signal is present, snapshots may be made. This means that E-W arrays like Westerbork and the ATCA are not particularly good for observations of these bodies, because their snapshot (u, v) coverage is one dimensional. Data can be combined from different days when the rotational aspects are similar as has been done by Leblanc *et al.* (1997) and de Pater (1980) in the case of Jupiter.

More resolution is not always better. As higher and higher resolution is obtained, the amount of flux density per beam is reduced. Eventually, the amount of flux density per beam gets smaller than the noise flux density per beam, and an image can no longer be made (with any meaning). The extreme case of this is the VLBA, which cannot be used to observe solar system bodies in thermal emission, because the lowest brightness temperature that it is sensitive to is of order 10^6 K. Before proposing to perform an observation of any solar system body, a careful calculation of the expected flux density per beam should be compared to the expected noise level (possibly in short snapshots) in order to investigate the feasibility of the observation. This is possible because the brightness temperatures of most solar system bodies are known (at least approximately).

2.3. Observing

Source tracking

Source motion has an important consequence for radio observations with a synthesis telescope because interferometric observations are typically referenced to a phase tracking center (see Lecture 2). To observe solar system objects, therefore, the phase tracking center must move with the object being observed. An accurate source position and its time derivative must therefore be obtained. For bodies with well known positions, the JPL "Horizons" system (Giorgini 1996, see also the web page at: http://ssd.jpl.nasa.gov/horizons.html) is a good place to get accurate ephemerides. Note, however, that each observatory has its own idiosyncrasies about how moving bodies are tracked, and how positions and their rates are specified to the telescope control system. Topocentric or geocentric positions and rates may be needed, they may need to be specified at a standard epoch (e.g. J2000) or for the epoch of date, and they may need to be referenced to any of a number of time frames (Coordinated Universal Time, International Atomic Time, Local Sidereal Time, etc.). At the VLA, the source position is specified on a source card and is generally referenced to the epoch of date. The time derivative of the source position and the horizontal parallax are specified on an accompanying "planetary motion" card. The source card specifies a stop time or duration of the observation in LST. However, the planetary motion card refers the time derivatives of the source position and the horizontal parallax to a time in IAT!

Spectral line observations of solar system objects must also account for their motion in some cases because such motion introduces a Doppler shift to spectral lines emitted by the object. If this shift is relatively constant over the duration of the observation (i.e., if the geocentric velocity of the observed body does not change significantly), then the receiving electronics simply need to be tuned to the Doppler-shifted line-center frequency. If this is not the case, then the change in line center with time must be accounted for in some manner. If it is possible to track the frequency shift during the observation, then an accurate estimate of the shift must be available beforehand. This is usually obtained from the same source as the positional ephemeris. If this is not possible, then the visibility data must be corrected for the time variable spectral shift after the observation.

Source flux density

The entire range of possible source flux densities may be found in observations of solar system bodies. At shorter wavelengths, the larger solar system bodies (Sun, Moon, and large planets) are the highest flux density sources in the sky. In the case of the VLA, the Sun is so strong that special hardware modifications must be utilized to perform solar observations (see section 4.). For very high flux density radio sources, the total system temperature of the antennas in an array used to observe the source has some contribution which is due to the source itself, the antenna temperature $T_{\rm ant}$. The Sun and Moon are extreme examples, where $T_{\rm ant}$ dominates $T_{\rm sys}$. For the planets, the problem is not quite so extreme, although $T_{\rm ant}$ may still contribute an appreciable fraction of the total system temperature. Let us examine an example. An observation of Venus is performed at 1.3 cm when the planet is at a distance of 0.5 AU. The antenna temperature due to the planet is given by:

$$T_{\rm ant} = \frac{\eta_a \, A_p \, V_o}{2 \, k_B} \,, \tag{31–1}$$

where η_a is the aperture efficiency of the antenna of physical area A_p, and V_o is the total flux density of the planet. Substituting for the total flux density (V_o) yields:

$$T_{\rm ant} = \eta_a \, A_p \, \frac{T_B}{\lambda^2} \, \frac{\pi \, R_v^2}{D_v^2} \,, \tag{31–2}$$

where T_B is the disk averaged brightness temperature of Venus at 1.3 cm ($T_B \sim$ 500 K), R_v is the radius of Venus (including the atmosphere, $R_v \sim$ 6120 km), and D_v is the distance to Venus (0.5 AU). For a 25 meter diameter antenna with an efficiency of 50%, the resulting antenna temperature is: $T_{\rm ant} \sim$ 15 K. Now, this is certainly not going to dominate the system temperature (the atmosphere at most locations provides at least this much emission), but when multiplied by the number of antennas in a large synthesis array, it may do so. This is the definition of "moderately strong" sources, i.e., the antenna temperature due to the total collecting area of the array is much greater than the system temperature (minus the contribution of the source) of any individual element of the array (Anantharamaiah *et al.* 1989, Kulkarni 1989).

2.4. Calibration issues

Most observations of solar system bodies are similar to other observations in terms of editing and calibration (Lecture 5). However, there can be some important differences as well. In the following sections we briefly discuss *only these differences* in flux density, amplitude, and phase calibration. The calibration of solar observations is considered separately later.

Flux density calibration

Generally, a source of known flux density (preferably *not* a planet!) is observed in order to fix the flux density scale. Getting the flux density scale right is important in observations of solar system bodies. This is because the information being extracted from the magnitude of the signals is related directly to the physical temperature and material properties of the body being observed. For solar system bodies, errors in the flux density scale which are of order 10% or larger produce undesirably large errors in these derived physical quantities.

Amplitude calibration

When the correlated flux density S_c on a given baseline becomes comparable with the total power $S_{\rm tot}$ or, equivalently, when the correlated brightness temperature T_C is comparable with $T_{\rm sys}$, the correlation coefficient is of order unity: $\rho = S_C/S_{\rm tot} = T_C/T_{\rm sys}$. The relationship between the correlation coefficient measured by a digital correlator and the true (analog) correlation coefficient is nonlinear. This nonlinearity is insignificant when the correlation coefficient is small, as it is for the vast bulk of VLA observing. For the VLA, therefore, a linear approximation to the "quantization correction" is performed by the on-line system (Butler 1998). However, for large correlation coefficients such as those encountered when observing solar system objects, the linear correction is not necessarily adequate and a non-linear correction is needed (e.g. Schwab 1979). These corrections are often referred to as the "Van Vleck correction." The task FILLM in AIPS (as of early 1998) can apply a much more accurate correction to archived VLA continuum data. VLA spectral line data are not corrected in this manner.

Phase calibration

It is possible that observations of a particular object were made with a poorly known ephemeris. While accurate positions and their time derivatives are readily available for the Sun, the Moon, the planets, and many asteroids, they are typically not available for comets at the time of an observation. If there is a positional error in right ascension ($\Delta\alpha$) or declination ($\Delta\delta$) when the body is observed, the visibilities are corrupted by an unknown, and perhaps time variable, phase term:

$$V'(u,v) = e^{i\Phi}\; V(u,v)\,, \tag{31–3}$$

where the phase term is given by:

$$\Phi = u\Delta\alpha\cos\delta + v\Delta\delta\,. \tag{31–4}$$

Once the ephemeris of a comet has been well determined, the visibility data may be corrected *post facto*. Conversely, if a relatively accurate ephemeris is used for the observations, then values of $\Delta\alpha$ and $\Delta\delta$ derived from the measured visibilities can be used to pinpoint positional errors in the ephemerides (Muhleman *et al.* 1985, Seidelmann *et al.* 1984).

Some observations of solar system bodies may involve distances that are close enough that the plane wave approximation of the incoming radiation breaks down. These are objects which are effectively in the near field of the interferometer. In this case, there is a phase term in the relationship between the visibility function and the sky brightness distribution which is often left out. For a two element interferometer, this phase term is given by (Thompson *et al.* 1991):

$$e^{2\pi\delta/\lambda}\,, \tag{31–5}$$

where δ is a geometrical quantity involving the physical positions of the two antennas forming the interferometer. Given projected physical locations of the two antennas as seen from the observed body (x_1, y_1) and (x_2, y_2), this quantity is given by:

$$\delta = \frac{x_1^2 + y_1^2 - (x_2^2 + y_2^2)}{2\,D}\,, \tag{31–6}$$

Table 31–2. Maximum Magnitude of Near-field Phase Term for $\lambda = 3.5$ cm

body	D (AU)	$\phi_{\rm max}$ (deg)
Mercury	0.6	22
Venus	0.3	44
Moon	0.00257	5150
NEA	0.03	440
Mars	0.5	27

where D is the distance to the observed body. To first order, the maximum value of this quantity reduces to:

$$\delta_{\rm max} \sim \frac{d_{\rm max}^2}{2\,D}\,, \tag{31–7}$$

where $d_{\rm max}$ is the maximum distance of any antenna in the synthesis array from the defined array center position. For the VLA, this maximum distance is about 20 km in the A configuration. This yields the maximum phase errors shown in Table 31–2 for observations of Mercury, Venus, the Moon, a near Earth asteroid, and Mars at a wavelength of ~ 3.5 cm. These errors are substantial, and must be accounted for in some way.

Currently, this phase term is taken into account automatically (by the on-line observing system) for observations of solar system objects at the VLA, but this has only been the case since about 1992. Observations prior to that need to be corrected for this term by adjusting the phases of the visibilities. Other observatories may or may not include this term, so caution should be exercised.

2.5. Imaging, deconvolution, and self-calibration

For observations which are not simple detection experiments, a high-fidelity, high-dynamic-range image of the brightness distribution (possibly as a function of wavelength) is often the desired goal. This is the case for most of the major planets, and many of the minor planets and satellites. These bodies generally have sufficient flux density that they can be self-calibrated. Therefore, in these cases, there is a general four-part cycle which results in the "best" image obtainable from the data: 1 - an initial model is obtained, possibly by fitting the (u, v) data; 2 - a deconvolution step, using that initial model, is performed to produce an image; 3 - the resultant image (or its component representation) is used to self-calibrate the (u, v) data; 4 - 1-3 are repeated to convergence (Lecture 10). Steps 1-3 will now be discussed in more detail.

The initial model

In all interferometric observations, there is a "hole" at the center of the (u, v) plane whose radius is defined by the shortest interferometer baseline. If the source being observed has significant flux density at large spatial scales, then the missing data on the baselines shorter than this hole radius can severely limit the quality of the image produced from observations of such a source. This problem is not unique to solar system observing, of course, and is discussed in

Lectures 8, 13, and 20. At the VLA, the larger planets, the Moon, and the Sun all suffer quite severely from this problem, because most of the flux density of these bodies is contained in the largest angular scales. Some of the difficulty presented by this problem may be overcome if a reasonable model of the source structure can be determined and specified. Since, in the case of the bodies which are not resolved by the primary beams of the antennas, we know the general shape of the sky brightness distribution (see Appendix A), a reasonable model may often be determined.

An initial model may be as simple as obtaining some estimate of the zero-spacing flux density (V_o) from any available short spacing visibility data, and using that with a uniform disk visibility function as the initial model. It may be as complicated as fitting the (u, v) data to some functional form, such as that shown in equation 31–34, combined with the positional offsets shown in equation 31–3. Or it may be something intermediate between these two extremes, such as fitting a uniform disk with no positional offsets. The choice of how complex to make the initial model is determined by the sensitivity of the data. It is senseless to fit complex models to data which are noisy or which are poorly sampled in the (u, v) plane. The (u, v) data fitting is usually accomplished by a least squares minimization of the residual between the visibilities and whatever model is being specified (Butler 1994, Lecture 16; the task OMFIT in AIPS). If at least one cycle of the model specification, deconvolution, and self-calibration has been completed, then fits to the image resulting from the deconvolution may actually be used, rather than fits to the (u, v) data (Grossman 1990).

In the case when the body is resolved heavily by the primary beams of the antennas (e.g. the Sun and Moon), specifying a model of this type is not so useful. In these cases, an actual measurement of V_o (from an independent single dish observation) may be used to constrain the total amount of flux density in the sky brightness distribution during the deconvolution step. This is similar to specifying an initial model, but is in fact superior because it utilizes a true measurement of V_o, rather than some fit or estimated value.

Deconvolution

Deconvolution generally employs one of the standard methods (e.g. CLEAN or MEM or their variants - Lecture 8), with two modifications: the specification of an initial model (or inclusion of a single dish observation); and the use of finite support (e.g. CLEAN windows). The specification of an initial model can be affected in several ways: for CLEAN based deconvolution (e.g. via IMAGR in AIPS) the model can be specified as a set of CLEAN components which are then subtracted from the (u, v) data as the first step of the CLEAN process (task CCMOD and modified variants in AIPS); for MEM based deconvolution (e.g. via VTESS in AIPS) the model can be specified as the "default" image; in either the MEM or CLEAN case, the model can be explicitly subtracted from the data (e.g. via UVSUB in AIPS) and the deconvolution performed on the resultant residual data set. In the subtraction case, the model must of course be added back in to the final image, after convolution with the proper restoring beam. The subtraction method works much better for MEM based deconvolution of planetary images. This is because MEM is based on the principle of maximum smoothness, and it is clear that there is a place in the images of planets which is not very smooth - the limb of the planet. In the residual data set, however, much of this edge is taken out, and the data can then be reliably deconvolved with MEM. Historically, CLEAN based methods have been preferred over MEM based methods for planetary images. One clear advantage of CLEAN based methods is the ability to specify regions of finite support (Lecture 8). For the planets, the extent of the sky brightness distribution is known quite well, because the locations and sizes of the planets are known. This is a great advantage in the deconvolution process (again, Lecture 8). There is an easy way of specifying this information in CLEAN based deconvolutions – through CLEAN windows. In the case of the planets, a CLEAN circle should be specified which is the size of the planet, plus about 1-2 synthesized beam widths. There is no straightforward way of specifying this type of information to MEM based methods (though the specification of the default image provides a part of this information).

Self-calibration

In order to correct for the effects of short timescale (less than the cycle time between the body of interest and the calibrator) variations in the atmosphere, self-calibration must be used. This is described in detail in Lecture 10. In general, self-calibration of data obtained while observing a solar system body is no different than self-calibration of any other data, but there is one detail which should be explicitly pointed out. Upon initial investigation, it might seem that the deconvolution step outlined above is unnecessary when data of high signal-to-noise ratio are obtained. Why not just fit the data, and use that specific fitted model to self-calibrate the visibilities? The answer is that a model which describes the full sky brightness distribution cannot be specified fully with only the few parameters available to fitting algorithms. The CLEAN components (or their equivalent) are the only good way to fully specify the model. For that reason, a deconvolution step should *always* be included before the self-calibration step. One note here is that historically, only the phases have been self-calibrated in planetary data. Since there is great concern over getting the flux density scale right, the feeling has been that an amplitude self-cal may corrupt the flux scale. This has not been shown in any rigorous manner, and needs to be investigated more completely.

An example

Figure 31–1 shows an example of the result of careful imaging of planetary interferometry data. This is an image of the thermal emission from the planet Mercury, made from data taken at the VLA at a frequency of about 8.5 GHz ($\sim$ 3.6 cm wavelength). This image (in combination with others at wavelengths from 3 mm to 20 cm) was used to obtain constraints on the material properties of the upper surface layer of Mercury, including the fact that it is much more transparent (at these wavelengths) than the lunar surface (Mitchell & de Pater 1994).

Confusion rejection

Although the motion of solar system bodies is a complication in the planning and observing stages, it can actually aid in rejecting confusion from background sources in interferometric observations. Since the background sources move relative to the phase center of the interferometer (which is centered on the body of interest), they move through the fringes of the interferometer response, essen-

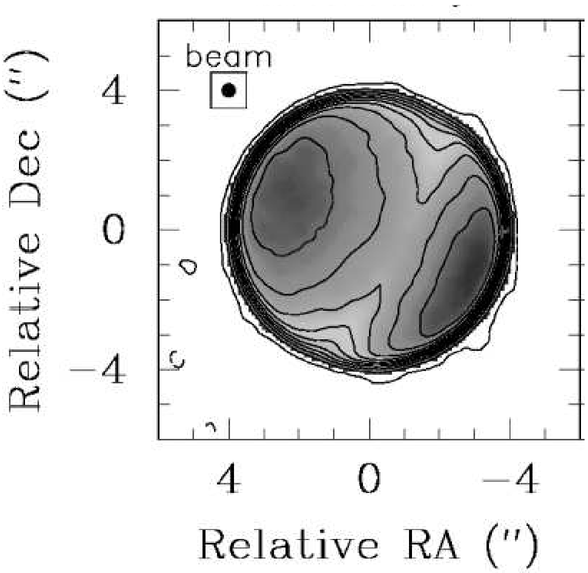


Figure 31–1. Thermal emission image of Mercury at 3.6 cm from the VLA (after Mitchell & de Pater 1994). Darker shades are higher temperatures. The "hot" and "cold" longitudes (resulting from Mercury's 3:2 resonance) are apparent.

tially averaging it down. This motion relative to the phase center also helps in determining where the confusing signals come from spatially. This works both for background sources and in cases where a body is observed relatively close to its primary (e.g. Mercury and the Sun, or Europa and Jupiter). In fact, the contribution of these confusing sources can be identified and nearly completely subtracted out in these types of observations (Sault & Noordam 1995, Thompson 1982, Briggs & Drake 1973, Briggs 1973).

Spectral line observing

When an atmospheric spectral line is being observed, the self-calibration cycle is usually performed using a data set which is some estimate of the thermal continuum emission. This is because in almost all cases, the thermal continuum is much stronger than the emission in the line. This thermal continuum emission may be estimated from spectral channels which are outside the line, or in some cases (where the line spans the entire measured bandwidth) may be just a combination of all of the spectral channels. As long as a bandpass calibration is used to calibrate the channel-to-channel variations, this is an acceptable approach.

As a final step before the analysis of the spectral line data, the continuum signal is generally either subtracted or divided out (Sault 1994, Lecture 12). In the case where there is no surface emission contribution present, the continuum is usually divided out. This provides a quantity (the line to continuum ratio) which is insensitive to errors in the absolute flux density calibration, and can be used in the subsequent data analysis (Gurwell *et al.* 1995). In the case where there is significant surface emission present, the line to continuum ratio loses its physical meaning, and in those cases, the continuum is usually subtracted out (Clancy *et al.* 1992).

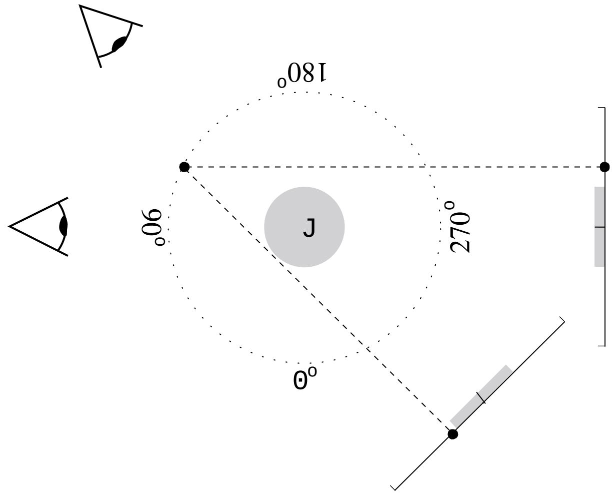


Figure 31–2. Geometry for multiple observations of Jupiter, allowing for 3-D reconstruction of the radiation belts. From Sault *et al.* (1997).

2.6. 3-D image reconstructions

Recently, a technique for reconstructing the full 3-D spatial structure of Jupiter's radiation belts (the regions where the energetic magnetospheric electrons reside) has been developed by Bob Sault and collaborators (Sault *et al.* 1997, Leblanc *et al.* 1997, de Pater & Sault 1998). The fundamental assumptions which allow for such a reconstruction are that the emission is optically thin and isotropic, and that the precise geometry of the Jovian system is known. The known geometry assumption is valid, since accurate ephemerides are available. The optically thin assumption is valid everywhere except where the planet itself blocks the emission from the portion of the radiation belts which is behind it. This is handled in an ad hoc (but seemingly effective) way in the reconstruction. The assumption of isotropic emission is clearly not valid, because the synchrotron emission is highly beamed (Legg & Westfold 1968).

Figure 31–2 shows a cartoon of the Jovian system, with two different observed central meridian longitudes illustrated. The 2-D projection of the radiation belts onto the plane of the sky for the two observed Central Meridian Longitudes (CMLs) results in a different projected l, m position for a given emission location in the 3-D radiation belts. Observations at many different geometries (both different CMLs, and different subearth latitudes) can be used to back out the information on the full 3-D structure of the belts. Specifically, each of the distinct observations of the sky brightness distribution is a 2-D projection of the full 3-D sky brightness distribution (in the optically thin case). Thus, the measured visibilities are samples of a 2-D slice through the full 3-D Fourier transform of the 3-D sky brightness distribution (the 3-D spatial coherence function - Sault *et al.* 1997, Bracewell 1995). So, if the visibilities from all of the different observations are gridded into the proper 3-D coordinates, they represent the 3-D Fourier transform of the full 3-D distribution in the radiation belts. Once the visibilities are gridded properly and Fourier transformed, a 3-D deconvolution can be used to recover a more accurate representation of the 3-D radiation belt structure.

There are certain problems with this reconstruction technique, like the isotropic emission assumption, the intrinsic time variability of the emission from the radiation belts, and the handling of the disk of Jupiter, which occludes the back portion of the belts. However, very nice images have resulted from the application of this technique to both ATCA and VLA data (Sault *et al.* 1997, Leblanc *et al.* 1997, de Pater & Sault 1998).

2.7. Conversion of units and coordinates

Flux density → brightness temperature

In general, it is the brightness temperature which is the interesting quantity in observations of solar system bodies, so flux densities are often converted into these units. For unresolved objects, or if only a few resolution elements are obtained on the object, then only crude information on the hemispherically averaged properties of the body can be obtained. In this case, the disk averaged brightness temperature can be calculated from:

$$T_B^d = V_0 \, \frac{\lambda^2}{2\,k_B} \, \frac{D^2}{\pi\,R_b^2} + T_{\rm cmb}\,, \tag{31–8}$$

where V_0 is the total flux density (the zero-spacing flux density) D is the distance to the body, R_b is the equivalent radius of the body, and $T_{\rm cmb}$ is the cosmic microwave background brightness temperature. The total flux density can be obtained by either examination of the visibilities (e.g. via the fitting described above), or by summing up the flux density in an image made from that data.

For higher resolution observations of solar system bodies, the observed flux density per beam can be converted to average brightness temperature over the beam (as a function of the sky coordinates l and m) via:

$$T_B^b(l,m) = F(l,m) \, \frac{\lambda^2}{2\,k_B} \, \frac{4\,\ln 2}{\pi\,B^2} + T_{\rm cmb}\,, \tag{31–9}$$

where $F(l,m)$ is the observed flux density per beam, and B is the diameter of the gaussian convolving kernel (e.g. the CLEAN restoring beam).

Sky coordinates → planetocentric cartographic coordinates

The desired final product in high resolution observations of the planets is often the sky brightness distribution (in whatever units it has been converted to) mapped onto a latitude and longitude grid for the object being observed. For this, a method of converting the sky coordinates (l and m) into latitude and longitude (ϕ and θ) coordinates on the body is required. Once such a method is available, interpolation into any of a number of different map projections is possible (e.g. Snyder 1987).

The coordinate conversion problem can be defined as: given sky coordinates l and m, equatorial and polar radii for the observed object R_e and R_p (true apparent radii, not projected apparent radii), and the subearth latitude and longitude ϕ_0 and θ_0, find the latitude and longitude corresponding to those sky coordinates. Note here that the body is assumed to be a prolate spheroid rather than the more general triaxial ellipsoid. Note also that the body-centric latitudes are used here, rather than body-graphic latitudes, but the conversion between the two is straightforward. Also, it is assumed in the following that the image has been rotated so that the projected polar axis is parallel to the m axis.

Consider a body-centered coordinate system with one axis through the north pole (the m' axis), one axis through the location of the 0° longitude line at 0° latitude (the n' axis), and the third axis at 90° longitude and 0° latitude (the l' axis). The body surface is defined by the points:

$$\frac{l'^2}{R_e^2} + \frac{m'^2}{R_p^2} + \frac{n'^2}{R_e^2} = 1\,. \tag{31–10}$$

The latitude and longitude of any point on that surface is given by:

$$\phi = \sin^{-1}\left(\frac{m'}{R'}\right) \qquad ; \qquad \theta = \tan^{-1}\left(\frac{n'}{l'}\right)\,, \tag{31–11}$$

where $R' = \sqrt{l'^2 + m'^2 + n'^2}$. The problem is then to convert the sky coordinates l, m, and n (note here that n is the axis with origin at the phase center and direction to the observer, which is somewhat different than the definition of the n axis in Lectures 2 and 19) into this body-centered coordinate system. This is achieved in two steps: 1 - n must be found; 2 - the sky coordinates must be rotated into the body-centered system.

The rotation of the sky coordinates l, m, and n into the body-centered coordinates l', m', and n' is achieved through two rotations: 1 - rotate about the l axis by the negative of the subearth latitude ($-\phi_0$); 2 - rotate about the m axis by the negative of the subearth longitude ($-\theta_0$). These two rotations yield:

$$\begin{aligned} l' &= l\cos\theta_0 - \sin\theta_0(n\cos\phi_0 - m\sin\phi_0) && (31–12)\\ m' &= m\cos\phi_0 + n\sin\phi_0 && (31–13)\\ n' &= l\sin\theta_0 + \cos\theta_0(n\cos\phi_0 - m\sin\phi_0)\,. && (31–14) \end{aligned}$$

To find the value of n, substitute the relations of equation 31–14 into equation 31–10 and solve for n. The result is a quadratic equation of the form:

$$A\,n^2 + B\,n + C = 0\,, \tag{31–15}$$

where A, B, and C are given by:

$$\begin{aligned} A &= \cos^2\phi_0 + b^2\sin^2\phi_0 && (31–16)\\ B &= 2m\cos\phi_0\sin\phi_0(b^2-1) && (31–17)\\ C &= l^2 + m^2(\sin^2\phi_0 + b^2\cos^2\phi_0) - R_e^2\,, && (31–18) \end{aligned}$$

with $b = R_e/R_p$. The solution to the above quadratic equation with $n > 0$ gives the value for n.

2.8. Data rectification to a common distance

It is often desirable to make observations of a body at one particular epoch resemble those at another epoch as much as possible, in order to compare results from different observations of the body. One obvious way to do this is to make the observations at one epoch appear as if they were undertaken when the body was at the distance that it was during the other epoch. In the extreme case, where the distance to the body is changing very rapidly, these two "epochs" may actually be during the same observing session, and in this case, all of the data (the visibilities) taken during that session need to have this adjustment applied, after some "standard" distance is chosen (usually the greatest distance during the session). In the more usual case, an entire data set has the same adjustment applied to it, in order to make the effective distance be what is desired. As an example, historically, observations of Jupiter have been referenced to the standard distance of 4.04 AU (Berge & Gulkis 1976). All observations of Jupiter should be "adjusted" to this standard distance in order to make intercomparisons between different data sets much easier.

To adjust observations taken when the body is at a distance D_1 to another distance D_0, the data must be modified in two ways. The first is a modification of the flux densities of the visibilities (the amplitudes). This modification is a simple multiplication of the visibility amplitude by the ratio of the distances squared:

$$|V'| = f^2 |V| \, , \tag{31–19}$$

where $f = D_1/D_0$. The second is a modification of the values of u and v for each of the visibilities. Because the projected spacing of the antennas scales linearly with distance, this modification is:

$$u' = \frac{u}{f} \tag{31–20}$$

$$v' = \frac{v}{f} \, . \tag{31–21}$$

This adjustment to the visibilities should be made after the initial calibration, but before any imaging and self-calibration.

3. Interferometric radar observations of solar system objects

Planetary radar astronomy traces its roots to the development of powerful radars during World War II. Radar research led to the development of a system which was capable of transmitting low frequency radar waves to the Moon and then detecting the returned echo (DeWitt & Stodola 1949). In the late 1950's, more than 15 years after the initial Moon success, radar systems sensitive enough to detect echoes from Venus were finally completed, and this is probably the real beginning of modern planetary radar astronomy (for a thorough history of planetary radar see Butrica 1996).

In single dish (monostatic) radar astronomy experiments, a nearly monochromatic wave is transmitted toward the target body for a time equal to the time it takes the wave to travel to the body and back. The transmitter is then turned off, and the reflected wave is received at the antenna. The received echo is spread

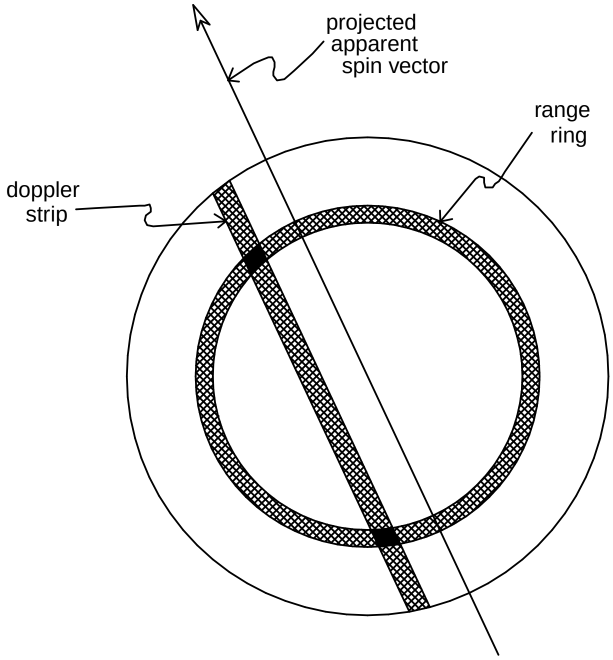


Figure 31–3. Projected hemisphere of a planetary radar target, showing a doppler strip and a range ring. The two solid areas are north-south ambiguous locations.

out in frequency, due to the different line of sight velocities of the different reflecting surface locations. This yields spatial resolution in 1 dimension - the direction perpendicular to the instantaneous projected apparent spin vector of the body. A spatial extent on the target bounded by a range of frequencies in the returned signal is termed a "Doppler strip". The simplest type of radar experiment utilizes only this 1 dimension of resolution (a so-called CW - or continuous wave experiment). In order to obtain resolution in the orthogonal direction, some sophisticated signal processing is introduced. This is fundamentally based on the principle that the time to reach different points on the target body (and return) is a function of the distance to the points. Locations which are near the limbs of the planets have round trip light times which are of course longer than the light times for locations near the subradar point. So, by encoding the outgoing signal with a specific time stamp, resolution in the direction along the line of sight is obtained. A spatial extent on the target bounded by a range of times of flight in the returned signal is termed a "range ring" or a "delay ring". So, two dimensional maps of the target can be obtained in experiments which have both time and frequency resolution - "delay-doppler" experiments. Note that Earth Synthetic Aperture Radar (SAR) observations (and those with the Magellan spacecraft) are based on this same principle.

Unfortunately, in planetary radar, there is an ambiguity in these two dimensional maps. Figure 31–3 illustrates this ambiguity, which arises because there are two locations (assuming the target is spherical) which contribute to the returned signal which is in a particular delay-doppler bin. Also, the delay-doppler technique has, until recently, been restricted to parts of the planet which are near the subradar point for rapidly rotating targets. Hagfors & Kofman (1991) and Hagfors & Tereshchenko (1991) describe this problem and two solutions to it (random long code, and frequency chirping). The ambiguity can be overcome if experiments with the target at many different observing geometries are available (Hagfors *et al.* 1968, Goldstein & Rumsey 1972), but this is not always possible. An alternative to varying the geometry of the target is varying the geometry of the receiving antenna. If several antennas at different locations are used to receive the range encoded signal, then two dimensional radar reflectivity maps without the ambiguity may be produced. This has been used to image parts of the surfaces of Mercury, Venus (Jurgens *et al.* 1980), and Mars (O'Brien *et al.* 1991). Another alternative is not to time encode the signal at all, but rather perform CW experiments. These experiments, in a number of geometries, and with some assumptions about the scattering behavior of the surface, allow for a reconstruction of the scattering properties of different surface locations (Hudson & Ostro 1990). A final alternative is to again not worry about the time encoding of the signal at all, but to receive the reflected signals at the antennas of an interferometric array, and use the techniques of synthesis imaging to make true (unambiguous) two dimensional radar maps. This technique has been used to image the planets Mercury (Butler *et al.* 1993), Venus (Haldemann *et al.* 1997), and Mars (Muhleman *et al.* 1991), and to obtain echoes from Titan (Muhleman *et al.* 1990). It is this technique that will be discussed here.

3.1. Telescopes, frequencies, and polarization

The only transmitter/synthesis array combination which has been used to date for this kind of planetary radar imaging is the Goldstone 70-m/VLA combination. The ATCA could also be used in experiments with Goldstone, but the time during which objects are visible from both sites is short, and the E-W configuration of the ATCA is not optimal. Goldstone has transmitters at both X-band (near 3.5 cm) and S-band (near 12.5 cm), but the VLA has no S-band receivers, so only X-band experiments can currently be performed. In the future, it is foreseen that the VLA antennas will be outfitted with S-band receivers, and in that case Goldstone/VLA experiments could be conducted at that wavelength. In addition, with Arecibo operating with more power at S-band, some Arecibo/VLA experiments are possible at that wavelength (but not all planetary bodies fall in the Arecibo beam). One final frequency possibility is that of Ka-band (near 33 GHz, or near 9 mm). There is also a plan to outfit the VLA antennas with receivers which operate at this frequency. The possibility also exists to place a transmitter at this frequency on the Goldstone 70-m antenna (in support of the Cassini mission). If that were done, then this would provide a nice short wavelength capability for this type of experiment.

Goldstone transmits nearly 500 kW of power at X-band, in one circular polarization. The VLA is used to receive both circular polarizations (and possibly the cross hand products). Traditionally, the received signal in the same circular polarization as that transmitted has been referred to as the SC (for Same-sense Circular) component. There are *many* other notations in the radar literature (including SS, unexpected, mismatched, depolarized), and this should be taken into account when reading the literature. The received signal in the opposite circular polarization as that transmitted has traditionally been called

Table 31–3. Total radar doppler spread of some solar system bodies.

body	ν_{dop} (Hz)
Mercury	400
Venus	60
Mars	27000
Titan & Galilean satellites	$\sim$ 1000
asteroids & comets	10's

the OC (for Opposite-sense Circular) component. Similarly, other notations exist (OS, expected, matched, polarized). SC and OC will be used in the following discussion.

3.2. Planning and scheduling

Planning and scheduling these experiments involve all of the complications mentioned above (notably, when to attempt such an experiment, and getting the ephemeris), in addition to some others. Because these experiments involve two observatories, coordinating the experiment is complicated. Time must be proposed for and allocated at Goldstone and the VLA for experiments which are expected to yield specific scientific results. For the Goldstone transmitter, time allocation is extremely competitive, and only a limited number of the joint Goldstone/VLA projects can be expected to be approved. One interesting complication for the Goldstone transmitter is that they must obtain permission from the FAA to transmit at low elevations to the west (because of air traffic around Los Angeles)!

Once an experiment has been approved, the characteristics of the receiving setup at the VLA must be chosen. These radar experiments are essentially spectral line experiments, where the "line" is the received echo. The transmitting frequency is constantly modified at the Goldstone transmitter so that the center of the line appears at a constant frequency as seen at the VLA (for X-band, this is currently 8510 MHz). The calculation of the necessary transmitting frequency to obtain this constant received frequency is complicated, and involves the precise JPL ephemeris (Goldstone is operated by JPL). The total expected width of the received echo can be calculated from the physical parameters of the observed body via:

$$\nu_{\text{dop}} = \frac{4\, R_b\, \omega\, \sin \phi_0}{\lambda}\,, \tag{31–22}$$

where ω is the rotation rate of the body. The resultant values for some solar system bodies are shown in Table 31–3 for 8510 MHz receiving frequency. The narrowest spectral channels which can be used in the correlator of the VLA are $\sim$ 380 Hz. This is multiplied by a factor of 2 (to 760 Hz) if correlations in two polarizations (e.g. RR & LL) are desired, and by another factor of two (to 1520 Hz) if all four Stokes correlations are desired. Comparison of these numbers with the total doppler spreads shown in Table 31–3 shows that care must be taken in choosing the combination of spectral width and Stokes correlations in some cases. If a receiving setup is selected which results in a spectral resolution which is wider than the total doppler spread of the body, then a penalty in SNR is incurred. This is because there are frequencies in the bandpass which contribute no signal, but still contribute noise (an optimum filter in the detection sense has a frequency width which is matched to that of the source). So, for example, when performing a joint Goldstone/VLA radar experiment to probe Mercury, a choice must be made regarding the tradeoff between SNR and polarization. If all 4 Stokes are desired, then a penalty of $\sqrt{2}$ is paid in SNR compared to the 2 Stokes case. In the Mars case, a spectral width must be chosen which is wide enough such that a location on the surface does not rotate through an entire channel width in the time chosen to make each snapshot. The snapshot time is determined by the rotation rate of locations near the subradar point, the distance to those locations, and the size of the synthesized beam. For these locations, the physical rate of motion is about 150 km in 10 minutes, which corresponds to about 0.4 arcseconds at 0.5 AU. The synthesized beam in the A configuration of the VLA at X-band is about this size, meaning that 10 minutes is about the right snapshot time. In 10 minutes, by moving 150 km, the change in frequency of a location near the subradar point is about 620 Hz (at 0.5 AU). So, spectral channels at least this wide must be chosen. A similar analysis must be done for each experiment of this type which is being proposed.

3.3. Data reduction (calibration, imaging, self-cal)

As in the case of continuum observations of solar system objects, the editing and calibration of radar spectral line data follows the same general steps as most "normal" VLA spectral line observations. One editing peculiarity is that there are often times during the course of an experiment when the transmitter fails for a period (sometimes for the whole experiment!). These times are recorded in a log at Goldstone, and they must of course be flagged in the data.

Once the initially calibrated data is in hand, the next step is the subtraction of the thermal continuum. Although in almost all cases the radar flux density dominates the received signal, there is always some component of thermal emission present. This thermal emission component must be subtracted out of the data to obtain the part of the received flux density due to the reflection of the radar wave. This presents certain problems with VLA data, since we are almost always operating in very narrow spectral channels. This means that the total bandwidth of the channels which can be combined together to estimate the thermal emission component (all channels which are outside of the expected radar echo - which may be all but the edge channels and the central channel) is usually quite narrow. So, the estimate of the thermal component may be quite noisy. With the planned VLA upgrade, this will not be a problem, as it will be possible to obtain both narrow spectral channels and a broad continuum simultaneously. Until that time, the noisy narrow band estimate is as good as it gets.

Once the thermal continuum has been subtracted, the next step is the iteration of self-calibration and imaging cycles to convergence. In the radar case, the OC data is usually self-calibrated, and those solutions applied to the SC data. This is because the OC flux density is so much stronger (for rock+soil surfaces, anyway) than the SC flux density, and also because the OC data is dominated by a point-like component near the phase center (the locations near the subradar point). This scheme will correct both polarizations for atmospheric errors, but will not correct the receiver errors of the SC data. If there is sufficient flux density in the SC polarization, then a final self-cal on that polarization may fix these errors. As in the case of the thermal continuum emission described above, only phase self-cal has been attempted on these data. The imaging cycles, as in the thermal continuum case above, usually involve an initial model fit to the visibilities. The functional form of the expected sky brightness distribution in the OC polarization can be written (Muhleman 1964):

$$I_{\rm oc}(\rho) = A_{\rm oc} \cos\theta_i \left(\frac{\alpha}{\sin\theta_i + \alpha\cos\theta_i}\right)^3 = A_{\rm oc}\sqrt{1-\rho^2}\left(\frac{\alpha}{\rho + \alpha\sqrt{1-\rho^2}}\right)^3 , \tag{31–23}$$

where $A_{\rm oc}$ is a scaling constant, and α is a measure of the roughness of the surface (it is the mean slope of surface facets). Hagfors (1964) provides an alternative functional form, but there is no *a priori* reason to prefer it over the one in equation 31–23 (see the discussion in Hagfors [1966] for comparisons between the two scattering functions). Unfortunately, when substituting equation 31–23 into equation 31–28, no analytical solution can be obtained. So, fitting the (u, v) data involves a time consuming least squares minimization where the above integral must be numerically calculated for every visibility point at every least squares iteration (Butler 1994).

3.4. Unit conversion

It is the radar reflectivity which is the interesting quantity in one of these joint radar experiments. If the target is unresolved by the synthesized beam of the receiving interferometer, then the returned flux density is related to the radar reflectivity via:

$$V_0 = \frac{P_t\, A_t}{4\,\pi\,\lambda^2\, D^2\, \Delta f}\, \frac{\pi\, R_b^2}{D^2}\, \sigma\, , \tag{31–24}$$

where P_t is the transmitted power, A_t is the effective transmitter area (which is a function of elevation), Δf is the frequency width of the spectral channels, and σ is the disk-averaged radar reflectivity. It is easy to see why planetary radar experiments are limited to relatively nearby objects - the received flux density is proportional to D^{-4}!

For experiments where the target is well resolved, the sky brightness distribution is related to the radar reflectivity via:

$$I(l, m) = \frac{P_t\, A_t}{4\,\pi\,\lambda^2\, D^2\, \Delta f}\, \frac{\pi\, B^2}{4\,\ln 2}\, \sigma(l, m)\, , \tag{31–25}$$

where B is the diameter of the convolving Gaussian (the CLEAN restoring beam), and $\sigma(l, m)$ is the average radar reflectivity over the region covered by the beam.

3.5. An example

Figure 31–4 shows an example of the result of a joint Goldstone/VLA radar experiment during the martian opposition of 1988. These images showed, for

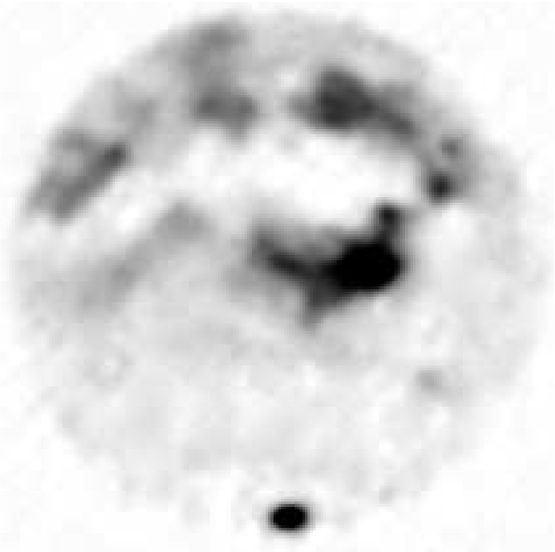


Figure 31–4. Snapshot image of the radar reflectivity of Mars obtained in a joint Goldstone/VLA radar experiment. Darker shades are higher reflectivity. The residual south polar ice cap is the spot at the bottom. The Tharsis volcanic region is prominent in the center of the image, as is the "Stealth" region, stretching to the west of Tharsis. From Butler (1994).

the first time, the incredibly reflective nature of the residual south polar ice cap, and the existence of a region which was nearly invisible to the radar which the observers dubbed "Stealth" (Muhleman *et al.* 1991).

4. Solar observing

Solar observations are performed in much the same way as observations of other solar system objects. The source is observed through a sequence of scans of some duration, interleaved with calibrator scans. The observing schedule corrects explicitly for horizontal parallax and the Sun's apparent motion. The Sun is large compared to the primary beam of the VLA antennas. Indeed, for all wavelengths shorter than 20 cm, the primary beam resolves the solar disk - the source fills the beam (and many sidelobes of the primary beam as well!). It is often the case that an observer wishes to track a particular feature on the Sun by tracking with the Sun's rotation. This is also accomplished in the observing schedule.

Due to its proximity, the Sun is a powerful radio source. So powerful, in fact, that it would completely saturate elements of the receiving system if the VLA were pointed at the Sun without hardware modifications. Such modifications are available. We discuss them only briefly here. Further detail can be found in Bastian (1989) or references therein.

4.1. Hardware modifications

The VLA was designed to observe faint cosmic sources for which, under most circumstances, receiver noise dominates the system temperature ($T_{\rm ant} \ll T_{\rm sys}$). When observing the Sun, this is no longer true - the Sun dominates the VLA system temperature at all bands, raising it by factors from several hundred to several thousand times the design values. Two fundamental hardware modifications are therefore necessary for solar observing.

First, it is necessary to reduce the effective gain of the receiving system of each antenna. This is accomplished by inserting attenuators in the front end of each antenna. The choice of the value for the insertion loss introduced by the switched attenuators is driven by constraints imposed by the "automatic level control" or "ALC" loops, and by the microwave emitting properties of the quiet and active Sun. Details can be found in Bastian (1989). The gain-reduction scheme implemented at the VLA is based on values of expected $T_{\rm ant}$ which are a compromise between active- and quiet-Sun observing. For all bands, *phase-constant* 20 dB switchable attenuators are used. That is, insertion and removal of the attenuators from the signal path introduce no more than a $1 - 2^\circ$ phase shift to the signal.

A second hardware modification is needed to calibrate the visibility amplitudes. Since, under the action of the ALC loop, the antenna gain is allowed to vary, the antenna gains must be monitored. Ordinarily, a switched noise signal of known amplitude is injected into each receiver input. Expressed in terms of an equivalent temperature, $T_{\rm cal}$, the amplitude of the injected signal is typically such that $T_{\rm cal} \sim 0.1 T_{\rm sys}$. Separate noise sources are provided for each band and for each of the two orthogonal senses of polarization. The $T_{\rm cal}$'s allow one to measure $T_{\rm sys}$. As the gain is inversely proportional to $T_{\rm sys}$, the antenna gains are corrected for variability via the so-called "$T_{\rm sys}$ correction".

When one observes the Sun, the system temperature is dominated by the Sun itself. The normal $T_{\rm cal}$'s are of no use because they are orders of magnitude smaller than $T_{\rm sys}$. Hence, when observing the Sun, special high-temperature $T_{\rm cal}$'s (*solar CALs*) are employed. Initially, because of the expense of retrofitting all twenty-eight VLA antennas with high-temperature noise sources, only four VLA antennas were provided with solar CALs. With ongoing upgrades of the antenna front ends, however, the 3.6 and the 20 cm bands have solar CALs on all antennas.

The usable dynamic range of the VLA with the above hardware modifications in place is restricted to roughly 13 dB, or a factor of twenty over quiet Sun conditions. If one is willing to suffer some degradation in accuracy of the $T_{\rm sys}$ measurement, then the factor that ultimately limits the dynamic range of the receiving system is saturation of a component—with this criterion, the dynamic range of the receiving system is roughly 20 dB, or a factor of 100 over quiet-Sun conditions. The available dynamic range in the ALC loop can therefore accommodate the increase in system temperature produced by most radio bursts.

4.2. Solar data calibration

As is the case for most other objects, solar visibility data are calibrated in phase by periodically observing a convenient phase calibrator. Phase calibrators are observed without the 20dB attenuators in place. Since the attenuators are phase-constant, the phase solution can be transferred from the calibrator to the source in the usual manner. Care has been taken to minimize phase-calibration errors through the use of phase-constant switched attenuators and optimizing the ALC operating point. Nevertheless, phase-calibration transfer errors of $\sim 5^\circ$ are to be expected routinely and transfer errors of $\sim 10^\circ$–15° are to be expected when the ALC operates far from its nominal regime, e.g., during a solar flare.

The visibility amplitudes cannot be referenced to any calibrator source, however, because few sidereal source can be detected when the 20 dB attenuators are in place. Instead, the flux scale is calibrated with the solar CALs and the visibility amplitudes are calibrated using the appropriate antenna-based $T_{\rm sys}$ corrections.

The total flux is related to the antenna temperature measured in a single polarization channel by $S_{\rm tot} = k_B T_{\rm ant}/\eta_a A$, where k_B is Boltzmann's constant, η_a is the antenna efficiency, and A is the antenna area. Since $T_{\rm sys} = T_{\rm ant}$, the online measurement of $T_{\rm sys}$ is sufficient to establish the total flux if η_a is known.

Three factors limit the accuracy of the solar amplitude calibration. First, the gains of the antennas that are not outfitted with solar CALs must be bootstrapped from those that have them, and for which $T_{\rm sys}$ is explicitly measured. To do this, we assume (D'Addario 1979): (i) that $T_{\rm ant} \gg T_{\rm sys}$; (ii) that the gain of each antenna is inversely proportional to $T_{\rm sys}$; and (iii) that $T_{\rm sys}$ is the same for all antennas for a given polarization. Assumption (i) is certainly true for all bands, as is assumption (ii) under most circumstances. Assumption (iii) is incorrect to the extent that antenna pointing errors result in variations in $T_{\rm sys}$ from one antenna to another, particularly when imaging a source near the solar limb.

A second difficulty involves the accuracy with which the solar CALs are known. Since there is no celestial source of known flux density bright enough to serve as a reference for the solar CALs, they must be measured individually in the field. The accuracy of this measurement is of order 20% (P. Lilie, private communication).

A third factor is the antenna efficiency. The antenna efficiency η_a is determined by many factors, including the quality of the reflector surface, aperture blockage, feed spillover, illumination taper, and diffraction. Each of these acts to block radiation or to scatter it into sidelobes. The efficiency of the main beam is therefore reduced. The angular size of the Sun is large compared to the primary beam for wavelengths $\lambda < 20$ cm at the VLA. Consequently many sidelobes of the primary beam fall on the source. As a result the effective efficiency is larger than it would be for a compact source in the main beam. The appropriate value of η_a is needed to measure the correct value of the total power $S_{\rm tot}$ collected by a given antenna. The amplitude of the correlation coefficient measured on a given baseline ij is the ratio of the correlated flux to the total flux, $\rho_{ij} = S_{ij}/S_{\rm tot}$, and the amplitude of the complex visibility is therefore $S_{ij} = S_{\rm tot}\rho_{ij} \propto \rho_{ij}/\eta_a$. Hence, the appropriate value of η_a is also needed to correctly calibrate visibility amplitudes.

In the case of the VLA, the appropriate value of η_a was determined using measurements on the moon by Bagri & Lilie (1993). The VLA antennas were pointed at the center of the lunar disk, which is very similar to the Sun in angular size, and careful measurements were made of the contribution of the moon to the system temperature. These have been compared with the brightness temperature of the moon at the appropriate lunar phase, averaged over the power pattern of a VLA antenna at the appropriate frequency. On the basis of comparisons with the COBE DMR moon-scanning calibration data, the absolute accuracy of the lunar brightness temperature is estimated to be ±3% (S. Keihm, private communication; see Bastian *et al.* 1996). At wavelengths of 1.3 and 2 cm, for example, these measurements yield effective antenna efficiencies of $\eta_a = 0.72 \pm 0.06$ and 0.73 ± 0.02, respectively. By way of comparison, the nominal values of the antenna efficiency are $\eta_a = 0.43$ and 0.52 at 1.3 and 2 cm.

It is important to point out that, for calibrating both the total flux and visibility amplitude scales even more accurately, additional corrections to the effective antenna efficiency are needed to account for 1) the fact that the antennas are often pointing at a target that is not on the Sun center – hence, the antennas sidelobes do not illuminate the Sun in a symmetric fashion – and/or 2) the details of the brightness distribution of the target itself.

4.3. Limitations

Imaging solar radio emission with an instrument like the VLA is limited by many factors. The dynamic range of solar images is typically of order 100:1 to 1000:1, often far below that achievable for other cosmic sources.

Calibration errors

Lecture 13 showed that the limitation to dynamic range imposed by calibration errors can be crudely estimated by considering a point source of unit amplitude. If an error in amplitude is expressed as ϵ and an error in phase as ϕ, the limitation to the dynamic range D due to either can be expressed as $\sqrt{MN}/x$ where x is either ϵ or ϕ, M is the number of independent measurements made, and N is the number of antennas. If we assume that $\epsilon \sim 20\%$ and $\phi \sim 10°$ for all antennas then $D \sim 100\sqrt{M/2}$. So even if no self-calibration is attempted, one might expect to obtain a dynamic range of several hundred to one, even in the presence of calibration errors. While calibration errors certainly contribute to limiting the dynamic range of solar images, they are often not the dominant factor.

Spatial confusion

For many solar observing programs, the single most important factor determining the dynamic range is *spatial confusion*. For a single antenna observation, confusion limits the instrumental sensitivity via the contributions to the system temperature from background sources in the beam. In synthesis imaging, the limitation to sensitivity imposed by confusion is the result of uncleanable sidelobe "clutter" in the image. While some observational strategies can minimize or eliminate the effects of confusing sources in the primary beam (Lecture 17) they are relevant only when the number of background sources is small. If the number of background sources is large, confusion becomes the major limitation to dynamic range.

The role of confusion in limiting the sensitivity of synthesis arrays has long been recognized at low frequencies, where the number of background sources in the primary beam is largest. Perley & Erickson (1984) have shown that if the baseline of a given interferometer is much larger than the element size (so that a large number of fringes cross the field of view), the interferometer response to a large number of point sources in the beam can be regarded as a random variable. It can then be shown that the confusion "noise", σ_c, is given by

$$\sigma_c = \rho \sqrt{\int B^2(\Omega)\, d\Omega \int S^2 n(S)\, dS}\,,$$

where B is the primary beam response, $n(S)$ is the source distribution function (i.e., the number of sources with flux densities between S and $S + dS$), and ρ is the r.m.s. fluctuation of the synthesized beam. In other words, the confusion "noise" may be expressed as an appropriately weighted, incoherent sum of the sidelobe responses to all sources in the primary beam. The confusion noise can be reduced to the extent that sidelobe responses to individual sources in the primary beam can be identified and removed.

(u, v) coverage

As noted in §2, the time variability of solar system objects, intrinsic or apparent, generally limits observations to a single observing session with a fixed array configuration. This means that the (u, v) coverage obtained is less extensive than that available from multiple observing sessions with multiple array configurations.

The difficulties presented by limited (u, v) coverage are twofold. The first problem, the "hole" at the center of the (u, v) plane whose radius is defined by the shortest interferometer baseline, is not specific to solar observing and is discussed in Lectures 8, 13, and 20. The problem can be particularly acute for solar imaging because most of the power emitted by the quiet Sun is in the largest angular scales. Depending on the requirements of a particular experiment three strategies are available at the VLA: i) ignoring the missing flux on short spacings, which is entirely justifiable for observations of compact transients; ii) measure and utilize the zero-spacing flux, which can be obtained from those antennas with solar CALs; iii) employ explicit measurements on the shortest spatial frequencies made by other instruments (e.g., a single dish). Maximum entropy image deconvolution algorithms (Lecture 8), provide a convenient way to introduce measurements made with more than one instrument to the image deconvolution problem.

A second problem with (u, v) coverage is closely related to that of confusion. Confusion is the major limitation to the dynamic range of solar images in many observing programs. We saw above that the confusion noise is proportional to the r.m.s. sidelobe level ρ of the synthesized beam. Improved (u, v) coverage reduces ρ and increases dynamic range. Unfortunately, the Sun's radio brightness distribution *varies* in time due to the Sun's rotation and intrinsic source variability, as we now discuss.

Source variability

The improved dynamic range that one might hope to obtain through improved (u, v) coverage may be sabotaged by time variability of the source. First, consider the Sun's rotation. Since the Sun's equatorial rotation velocity is

~ 2 km s^{-1}, a source at the center of the Sun's disk (e.g., an active region) will apparently move away from the center of the disk at 9.3 arcsec hr^{-1}. On the other hand, since the Sun is spherical, sources near the limb apparently move very little due to rotation.

The success with which one can eliminate the effects of solar rotation depends on the type of imaging program. If the field of interest is small compared to the radius of the Sun, one may choose to explicitly correct for the rotation of the Sun by tracking a particular feature. As the field becomes large, however, parts of the image will be smeared and/or distorted by differential motion across the field. Ideally, the region should be imaged in a time that is short compared to that for significant degradation of the image by the Sun's rotation.

We now turn to the problem of intrinsic variability in the radio brightness distribution. Consider a discrete source at some location (x, y) with brightness S Jy/beam. Let its lifetime be τ but suppose the observation is of duration $T > \tau$. Then while it is true that, limitations imposed by confusion aside for the moment, the r.m.s. fluctuations on the image are reduced by $\sqrt{T/\tau}$ relative to those expected from an integration of duration τ, it is also true that the brightness at (x, y) is averaged down to $\tau S/T$. The net signal-to-noise ratio at (x, y) is thus reduced by $\sqrt{\tau/T}$. In other words, the temporal smearing of short-lived contributions to the Sun's radio brightness distribution may lead to a net reduction in effective dynamic range.

A further complication is that if the radio brightness distribution on the sky varies as a function of time, due to either solar rotation or intrinsic variability, the convolution relation between the brightness distribution and the instrument response function is no longer valid. The situation is further exacerbated by the fact that the temporal variability in turn depends on position. The consequent breakdown of deconvolution algorithms limits the final dynamic range. Hence, in addition to spatial confusion, temporal confusion limits the dynamic range of solar images.

What is required, then, is improved (u, v) coverage over time scales that are short in comparison to the lifetime of the phenomenon of interest, or compared to the time for solar rotation to have a significant effect. Frequency synthesis is one possibility (Lecture 21). Ultimately, however, both the problem of (u, v) coverage and the problem of source variability must be addressed with an array containing many more antennas than the VLA. The radioheliograph at Nobeyama, Japan, (84 antennas) is such an array (Nakajima *et al.* 1994) .

4.4. An example

An example of an observations of a solar flare is shown in Figure 31–5. Solar radio bursts are extremely dynamic, varying in intensity, morphology, and polarization on short time scales. This observation was made by the VLA (C configuration) on 17 June 1989. It illustrates both the capabilities and the shortcomings of the VLA as a solar imaging instrument. The array was divided into two subarrays to enable simultaneous imaging in the 14.9 GHz (2 cm) and 4.9 GHz (6 cm) bands. The source size and structure was a good match to the C configuration in the 6 cm band. Unfortunately, the source was strongly

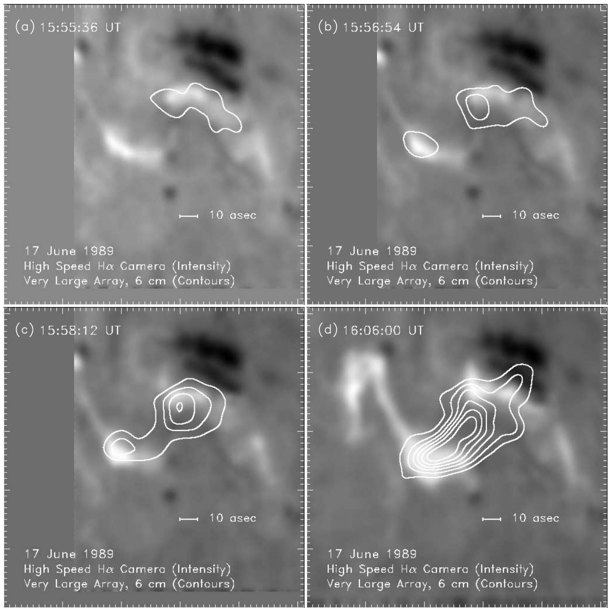

(a) 15:55:36 UT
10 asec
17 June 1989
High Speed Hα Camera (Intensity)
Very Large Array, 6 cm (Contours)
(b) 15:56:54 UT
10 asec
17 June 1989
High Speed Hα Camera (Intensity)
Very Large Array, 6 cm (Contours)
(c) 15:58:12 UT
10 asec
17 June 1989
High Speed Hα Camera (Intensity)
Very Large Array, 6 cm (Contours)
(d) 16:06:00 UT
10 asec
17 June 1989
High Speed Hα Camera (Intensity)
Very Large Array, 6 cm (Contours)

Figure 31–5. Example of a flare observed by the VLA. The contours represent the 6 cm emission while the grayscale shows the corresponding Hα emission. Large sunspots are seen to the NW. (from Bastian & Kiplinger 1989).

overresolved in the 2 cm band for most of the flare and the 2 cm data were therefore of limited use.

Appendix A

Expected Visibility Functions

All solar system bodies with radii greater than about 200 km have a shape which can be described fairly well as a triaxial ellipsoid. Since these larger bodies are the ones which can be imaged by current interferometric arrays, examining the expected interferometer response when such a body is observed is informative. From Lecture 2 the 2-D Fourier transform relationship between the spatial coherence function (the visibility function) and the sky brightness distribution is:

$$V(u,v) = \iint \mathcal{A}(l,m)\ I(l,m)\ e^{-2\pi i(ul+vm)}\ dl\,dm\,. \tag{31–26}$$

Now, consider any sky brightness distribution which is circularly symmetric, and assume that the primary beam response is also circularly symmetric. In this case, the 2-D Fourier transform relationship reduces to a 1-D Hankel transform (of zero order) relationship (Bracewell 1986):

$$V(q) = 2\pi \int \mathcal{A}(r)\ I(r)\ J_0(2\pi rq)\ r\ dr\,, \tag{31–27}$$

where $r = \sqrt{l^2 + m^2}$ is the radial image plane coordinate, $q = \sqrt{u^2 + v^2}$ is the radial (u, v) plane coordinate, and J_0 is the Bessel function of the first kind of order zero. If the circularly symmetric sky brightness distribution is bounded at some maximum apparent radius R (as is the case here), then the visibility function becomes:

$$V(\beta) = 2\pi R^2 \int_0^1 \mathcal{A}(\rho)\ I(\rho)\ J_0(2\pi\rho\beta)\ \rho\ d\rho\,, \tag{31–28}$$

where $\rho = r/R$ is the normalized radial image plane coordinate, and $\beta = Rq$ is the apparent radial (u, v) plane coordinate.

Even if the source is not circularly symmetric, but is elliptically symmetric, then this same visibility function results, just with a more complicated definition for β. All that is necessary is a coordinate transformation from the original l, m coordinates to a new set in which the source *is* circularly symmetric. For an elliptical source with the long axis of apparent half-length R making an angle Θ with the l axis (measured counterclockwise), β can be redefined as:

$$\beta = R\sqrt{v'^2 + u'^2}\,, \tag{31–29}$$

where

$$u' = u\ \cos\Theta - v\ \sin\Theta\,, \tag{31–30}$$

and

$$v' = b\ (u\ \sin\Theta + v\ \cos\Theta)\ , \tag{31–31}$$

with b the ratio of the long axis to the short axis. This is equivalent to a rotation to make the polar axis coincide with the m axis, followed by a scaling of the m coordinate. There are other ways of specifically defining β (e.g. Rudy 1987, Briggs & Sackett 1989), but the result is equivalent.

As the simplest case for the sky brightness distribution, assume that it is some constant value, I_0 (often referred to as a "uniform disk"). In this case, the visibility function reduces to (ignoring the primary beam response):

$$V(\beta) = I_0 \pi R^2\, \frac{J_1(2\pi\beta)}{\pi\,\beta}\,. \tag{31–32}$$

This is a good first order model for the radio emission from planetary surfaces and atmospheres. However, modeling the radio emission more precisely shows that the emission toward the limbs of the planets is expected to be reduced from that near the center - this is called "limb darkening". If the limb darkening is parameterized in the following way:

$$I(\rho) = I_0\ \cos^p(\theta_i) = I_0\,(1 - \rho^2)^{p/2}\,, \tag{31–33}$$

where θ_i is the look angle (angle between line of sight and surface normal on the planet), and p is a measure of the magnitude of the limb darkening, then the visibility function reduces to (again, ignoring the primary beam response):

$$V(\beta) = I_0 \pi R^2\, \Lambda_q(2\pi\beta)\,, \tag{31–34}$$

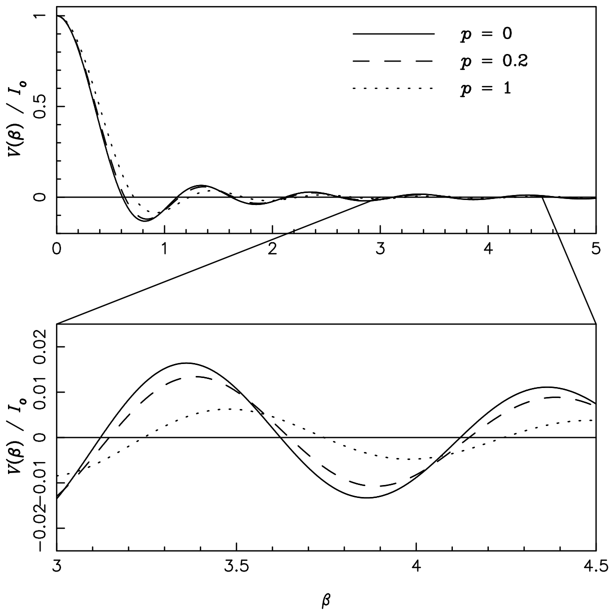


Figure 31–6. Expected visibility function for several values of the limb darkening parameter p. Inset shows detail at larger values of β.

where $q = 1 + p/2$, and $\Lambda_q(z)$ is the Lambda function of order q evaluated with argument z:

$$\Lambda_q(z) = \Gamma(q+1)\left(\frac{1}{2}z\right)^{-q} J_q(z)\,. \qquad (31\text{–}35)$$

Equation 31–34 reduces to the relation in equation 31–32 for a uniform disk ($p = 0$). Figure 31–6 shows a plot of the visibility function for several values of the limb darkening parameter p. As p is increased, the nulls of the visibility function occur at larger values of β, i.e., the planet looks effectively smaller than it truly is. Note that in reality, the thermal emission from planetary surfaces and atmospheres does not really follow in detail this specific parameterization of the sky brightness distribution, it is merely a way of parameterizing the distribution so that the integral equation for the visibility function can be analytically solved. Figure 31–7 shows the real and imaginary part of visibility samples from an actual observation of Venus with the VLA at 1.3 cm. Also shown in that figure is the fit to the real part of these visibilities, with $p = 0.18$. The fit is so good that you can't see the line of the fit because it is covered up by the visibility plot points. Note that the visibility phase alternates between 0 and $\pm\pi$ from lobe to lobe of the visibility function.

In the case of emission from a planetary surface, there is a polarized component. This is because the passage of the E-M wave from the surface into the atmosphere (or free space) polarizes the initially unpolarized radiation (Heiles & Drake 1963). In this case, we can also derive the expected response of the interferometer to this polarized emission. For a dielectric sphere, consider the difference between the brightness temperature in the polarization directions perpendicular and parallel to the plane of incidence (the plane formed by the surface

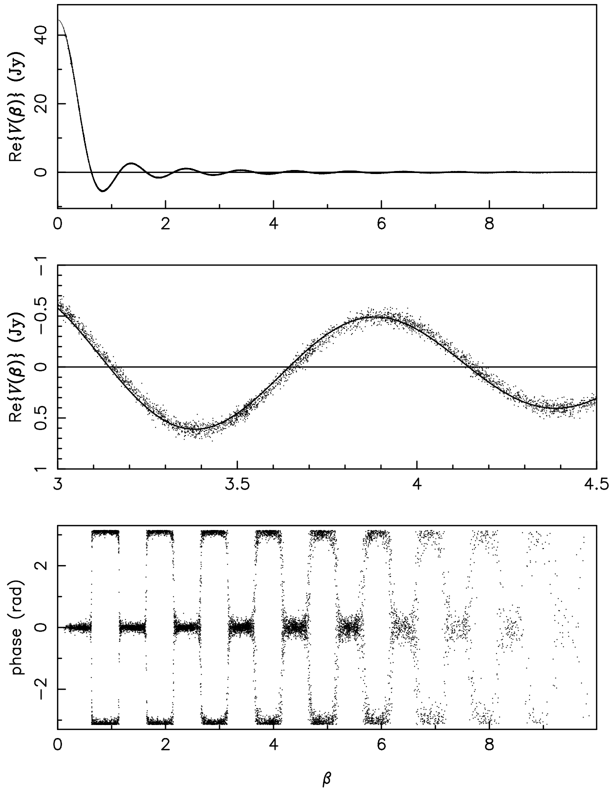


Figure 31–7. Measured visibilities from a VLA observation of Venus at 2 cm (discrete points), as well as model fit (solid line). Top panel is the real part of the visibility, middle panel is a blowup of the region $3.0 < \beta < 4.5$, bottom panel is visibility phase. The fit in the top two panels has $p = 0.18$. Only 2% of the visibility data samples from the experiment are actually plotted (every 50^{th} visibility).

normal and the line of sight direction). The brightness temperatures in these two polarization directions are different because of the difference in the Fresnel transmission coefficients in these two directions. In this case, as in the total emission case, the 2-D Fourier relationship between the visibility function and the sky brightness distribution can be reduced to a 1-D transform. If that result is normalized by the zero-spacing flux density ($V(\beta = 0) = V_0$), then this normalized polarized visibility function is given by another Hankel transform (of order two) (Rudy *et al.* 1987):

$$V_p(\beta) = \int_0^1 \mathcal{A}(\rho)\ (R_{||} - R_{\perp})\ J_2(2\pi\rho\beta)\ \rho\ d\rho\,, \tag{31–36}$$

where $R_{||}$ and $R_{\perp}$ are the Fresnel reflection coefficients in the parallel and perpendicular directions. For dielectric contrast at the interface of ϵ, and for incidence angle $\theta_i = \sin^{-1}\rho$, these reflection coefficients are:

$$R_{||}(\theta_i) = \left[\frac{\epsilon\cos\theta_i - \sqrt{\epsilon - \sin^2\theta_i}}{\epsilon\cos\theta_i + \sqrt{\epsilon - \sin^2\theta_i}}\right]^2 \tag{31–37}$$

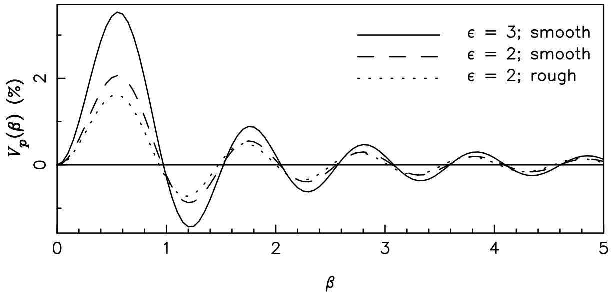


Figure 31–8. Expected polarized visibility function for two values of the dielectric constant. The rough curve is calculated for a surface with an exponential distribution of surface slopes, with a mean slope of $\alpha = 0.4$ (Muhleman 1964).

$$R_\perp(\theta_i) = \left[\frac{\cos\theta_i - \sqrt{\epsilon - \sin^2\theta_i}}{\cos\theta_i + \sqrt{\epsilon - \sin^2\theta_i}}\right]^2 . \tag{31–38}$$

This representation of the equivalent polarized response has no analytical solution, and must be numerically integrated. Note that the above expression is only formally valid for a perfectly smooth surface. The roughness of planetary surfaces causes a modification of the response (Golden 1979, Hagfors & Moriello 1965). An example of the expected polarized response is shown in Figure 31–8 for two values of the dielectric constant, along with one for a rough surface.

For antennas with circular feeds, and hence measured visibilities in Stokes RL and LR (V_{LR}, and V_{RL} - Lecture 6), the quantity V_p is calculated via:

$$V_p = \frac{\mathrm{Re}\left\{V_{RL} + V_{LR}\right\}\cos 2\psi + \mathrm{Im}\left\{V_{RL} - V_{LR}\right\}\sin 2\psi}{V_0} , \tag{31–39}$$

where ψ is the angle the baseline makes with the east-west direction ($\psi = \tan^{-1}(u/v)$).

In the above derivations, the primary beam response was assumed to be constant across the region of interest for the final steps. This is not always the case in solar system observations, because the planets, Sun, and the comae of comets can be quite large in the sky. In fact, except at very long wavelengths, the Sun and Moon are much larger than the extent of the main lobe of the primary beam for almost all currently operating synthesis arrays. In this case, the expected response will of course be modified by the primary beam response, and the results shown above will not be correct. From Table 31–1, it can be seen that in addition to the Sun and Moon, comet atmospheres are heavily resolved by the primary beams of large radio telescopes. Jupiter and Venus are also partially resolved by the primary beams of large radio telescopes which are operated at the shorter wavelengths, and by most millimeter telescopes. For comparison, the FWHM of a 25 meter diameter antennas at 22 GHz is about 130 arcseconds.

References

Anantharamaiah, K. R., Ekers, R. D., Radhakrishnan, V., Cornwell, T. J., & Goss, W. M. 1989. in *Synthesis Imaging in Radio Astronomy* eds. R. A. Perley, F. R. Schwab, & A. H. Bridle (San Francisco: PASP), 431–442.

Bagri, D., & Lilie, P. 1993. VLA Test Memo No. 170, NRAO.

Bastian, T. S. 1989. in *Synthesis Imaging in Radio Astronomy* eds. R. A. Perley, F. R. Schwab, & A. H. Bridle (San Francisco: PASP), 395–413.

Bastian, T. S., & Kiplinger, A. L. 1991. *Max '91 Workshop #3*, Estes Park, CO, ed. R. Winglee, A. Kiplinger, 153–162.

Bastian T. S., Dulk, G. A., & Leblanc, Y. 1996. *ApJ*, 473, 539–549.

Berge, G. L., & Gulkis, S. 1976. In *Jupiter*, ed. T. Gehrels, Tucson: Univ. Ariz. Press, 621–692.

Bracewell, R. N. 1986. *The Fourier Transform and Its Applications, 2nd Ed., rev.*, New York: McGraw-Hill.

Bracewell, R. N. 1995. *Two-dimensional imaging*, Inglewood Cliffs: Prentice-Hall.

Briggs, F. H., & Drake, F. D. 1973. *ApJ*, 182, 601–607.

Briggs, F. H. 1973. *ApJ*, 182, 999–1011.

Briggs, F. H., & Sackett, P. D. 1989. *Icarus*, 80, 77–103.

Butler, B. J. 1998. VLA Test Memorandum No. 212, NRAO.

Butler, B. J. 1994. *Ph.D. Thesis*, Caltech.

Butler, B. J., Muhleman, D. O., & Slade, M. A. 1993. *J. Geophys. Res.*, 98, 15003–15023.

Butrica, A. J. 1996. *To See the Unseen: A History of Planetary Radar Astronomy*, Washington D.C., NASA.

Christiansen, W. N., & Warburton, J. A. 1955. *Aust. J. Phys.*, 8, 474–486.

Clancy, R. T., Grossman, A. W., & Muhleman, D. O. 1992. *Icarus*, 100, 48–59

D'Addario, L. R. 1979, VLA Scientific Memo No. 130, NRAO.

de Pater, I. 1980. *A&A*, 88, 175–183.

de Pater, I., & Sault, R. J. 1998. *J. Geophys. Res.*, 103, 19973–19984.

DeWitt, J. H., & Stodola, E. K. 1949. *Proc. IRE*, 37, 229–242.

Giorgini, J. D., Yeomans, D. K., Chamberlin, A. B., Chodas, P. W., Jacobson, R. A., Keesey, M. S., Lieske, J. H., Ostro, S. J., Standish, E. M., Wimberly, R. N. 1996. *BAAS*, 28, 1158.

Golden, L. M. 1979. *Icarus*, 38, 451–455.

Goldstein, R. M., & Rumsey, H. C. 1972. *Icarus*, 17, 699–703.

Grossman, A. W. 1990. *Ph.D. Thesis*, Caltech.

Gurwell, M. A., Muhleman, D. O., Shah, K. P., Berge, G. L., Rudy, D. J., & Grossman, A. W. 1995. *Icarus*, 115, 141–158.

Hagfors, T., & Tereshchenko, E. 1991. *Radio Sci.*, 26, 1199–1203.

Hagfors, T., & Kofman, W. 1991. *Radio Sci.*, 26, 403–416.

Hagfors, T., Nanni, B, & Stone, K. 1968. *Radio Sci.*, 3, 491–509.

Hagfors, T. 1966. *J. Geophys. Res.*, 71, 379–383.

Hagfors, T., & Moriello, J. 1965, *Radio Sci.*, 69D, 1614–1615.

Hagfors, T. 1964. *J. Geophys. Res.*, 69, 3779–3784.

Haldemann, A. F. C., Muhleman, D. O., Butler, B. J., & Slade, M. A. 1997. *Icarus*, 128, 398–415.

Heiles, C. E., & Drake, F. D. 1963. *Icarus*, 2, 281–292.

Hudson, R. S., & Ostro, S. J. 1990. *J. Geophys. Res.*, 95, 10947–10963.

Jurgens, R. G., Goldstein, R. M., Rumsey, H. R., & Green, R. R. 1980. *J. Geophys. Res.*, 85, 8282–8294.

Kulkarni, S. R. 1989. *AJ*, 98, 1112–1130.

Leblanc, Y., Dulk, G. A., Sault, R. J., & Hunstead, R. W. 1997. *A&A*, 319, 274–281.

Legg, M. P. C., & Westfold, K. C. 1968. *ApJ*, 154, 499–514.

McCready, L. L., Pawsey, J. L., & Payne-Scott, R. 1947, *Proc. Roy. Soc. A*, 190, 357–375.

Mitchell, D. L., & de Pater, I. 1994, *Icarus*, 110, 2–32.

Muhleman, D. O., Butler, B. J., Grossman, A. W., & Slade, M. A. 1991. *Science*, 253, 1508–1513.

Muhleman, D. O., Grossman, A. W., Butler, B. J., & Slade, M. A. 1990. *Science*, 248, 975–980.

Muhleman, D. O., Berge, G. L., Rudy, D. J., Niell, A. E., Linfield, R. P., & Standish, E. M. 1985. *Celest. Mech.*, 37, 329–337.

Muhleman, D. O. 1964. *AJ*, 69, 34–41.

Nakajima, H., Enome, S., Shibasaki, S., Nishio, M., Takano, T., *et al.* , 1994. *Proc. IEEE*, 82, 705–713.

O'Brien, T. C., Jurgens, R. F., Slade, M. A., Howard, S. D., Moore, H. J., & Thompson, T. W. 1991. *LPSC XXII*, 991–992.

Perley, R. A., & Erickson, W. C. 1984. VLA Scientific Memo No. 146, NRAO.

Rudy, D. J., Muhleman, D. O., Berge, G. L., Jakosky, B. M., & Christensen, P. R. 1987, *Icarus*, 71, 159–177.

Rudy, D. J. 1987. *Ph.D. Thesis*, Caltech.

Sault, R. J. 1994. *A&AS*, 107, 55–69.

Sault, R. J., & Noordam, J. E. 1995. *A&AS*, 109, 593–595.

Sault, R. J., Oosterloo, T., Dulk, G. A., & Leblanc, Y. 1997. *A&AS*, 324, 1190–1196.

Schwab, F. R. 1979. VLA Computer Memorandum No. 150, NRAO.

Seidelmann, P. K., Kaplan, G. H., Johnston, K. J., & Wade, C. M. 1984, *Celest. Mech.*, 34, 39–48.

Snyder, J. P. 1987. Professional Paper 1395, USGS.

Thompson, A. R. 1982. *IEEE Trans. Ant. Prop.*, 30, 450–456.

Thompson, A. R., Moran, J. M., & Swenson, G. W. Jr. 1991. *Interferometry and Synthesis Imaging in Radio Astronomy, 2nd ed.*, Krieger Publishing Co., Malabar, 1991

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

32. The Hamaker-Bregman-Sault Measurement Equation

R. J. Sault
Australia Telescope National Facility, Epping, NSW 2121, Australia

T. J. Cornwell
National Radio Astronomy Observatory, Socorro, NM 87801, U.S.A.

Abstract. This lecture describes the use of the formalism of Jones and Mueller matrices in describing the polarimetric response of a radio interferometer. This formalism in radio interferometry has become known as the "Hamaker-Bregman-Sault measurement equation". Apart from providing a compact description, this formalism allows general results about the limits of polarimetric calibration and self-calibration to be determined relatively easily. It also provides an algorithmically general approach to the calibration of an interferometer, and has been used in the implementation of calibration in AIPS++.

1. Introduction

The earliest detections of polarized emission at radio wavelengths were made from the Sun in 1946 (Pawsey & Bracewell 1955) and from the Crab Nebula in 1956 (Mayer, McCullough & Sloanaker 1957). Although these early observations are some 40-50 years old, optical polarimetry is a far older field. It is not surprising, then, that a rich set of polarimetric formalisms have been developed by the optical community. These include Stokes parameters to describe the polarization state of light, and Jones and Mueller matrices to describe the transformation of the polarization state by the propagation medium and the detecting instrument. Stokes parameters are now almost universally used in astronomy. Although single-dish radio astronomers occasionally express their polarimetric calibration in terms of Mueller and Jones matrices (e.g. Turlo et al. 1985), the technique is just as applicable to radio interferometry.

This lecture consists of three parts. Firstly, it introduces the formalism of Jones and Mueller matrices as they apply to radio interferometry. This is based on the paper by Hamaker, Bregman and Sault (1996), which elaborated the use of this formalism in radio interferometry. The application of this formalism to radio interferometry has become known as the Hamaker-Bregman-Sault (or HBS) measurement equation. Other more general introductions to these concepts are given by Hecht & Zajac (1974) and Tinbergen (1996). However, the links with the optical community are important not only for providing a compact formalism: the second part addresses the level of polarimetric calibration that can be achieved with various calibrator observations. This section is based on the paper by Sault, Hamaker & Bregman (1996). Finally, this lecture discusses an algorithmic approach to polarimetric imaging and how it is implemented in AIPS++ (see Cornwell & Wieringa 1997).

2. Jones Matrices

A convenient way to represent an electric field of a quasi-monochromatic wave or voltage, $E_0 \cos(\omega t + \phi)$, is in terms of its so-called "complex amplitude" or phasor, $E = E_0 \exp(i\phi)$. This representation embodies the amplitude and phase of the wave. As we are usually only interested in linear systems[1], the net result of a system on a wave or voltage is to multiply its complex amplitude by a complex gain term. When polarization is important, two complex amplitudes, of orthogonal polarization states, are needed to describe a wave. Although orthogonal linear polarization states (x and y) would seem the obvious states to choose (and are chosen in most optical polarimetry texts), orthogonal circular states are equally reasonable choices (note the word "linear" in "linear system" and "linear polarization" refers to two distinct concepts). Because of the greater prevalence of circularly polarized feeds in radio astronomy, orthogonal circular states will be used in this lecture. Note that this choice is simply that of picking a convenient coordinate system, and has no fundamental importance.

Even for a linear system, some polarization conversion from one state to another can occur. We can describe such systems by a (potentially) complex-valued 2×2 matrix. If $E_{\rm R}$ and $E_{\rm L}$ are the complex amplitudes of the right and left circular waves/voltages, then the output of a system is given by

$$\begin{pmatrix} E'_{\rm R} \\ E'_{\rm L} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} E_{\rm R} \\ E_{\rm L} \end{pmatrix} \tag{32–1}$$

That is, the output polarization states are some linear combination of the input polarization states. This 2×2 response matrix is known in polarimetry literature as a Jones matrix, after R.C. Jones, who formulated their use in 1941.

Jones matrices can be readily derived for simple systems, and the result of passing a complex system can be expressed as the matrix product of the Jones matrices of the component subsystems

$$\mathbf{J}_{\rm overall} = \mathbf{J}_1 \mathbf{J}_2 \mathbf{J}_3 \tag{32–2}$$

etc. In our context, a Jones matrix can represent the effect of the propagation medium (e.g. ionosphere or troposphere), the antenna feeds or part of the electronic signal path up to the correlator. The most common Jones matrices are "antenna gain"

$$\mathbf{J}_{\rm gain} = \begin{pmatrix} g_{\rm R} & 0 \\ 0 & g_{\rm L} \end{pmatrix}, \tag{32–3}$$

polarization leakage (the so-called "D terms")

$$\mathbf{J}_{\rm leakage} = \begin{pmatrix} 1 & D_{\rm R} \\ -D_{\rm L} & 1 \end{pmatrix}, \tag{32–4}$$

and rotation. When using orthogonal-circular states as the representation, the rotation Jones matrix is

$$\mathbf{J}_{\rm rotation} = \begin{pmatrix} \exp(-i\theta) & 0 \\ 0 & \exp(i\theta) \end{pmatrix}, \tag{32–5}$$

[1] The signal path in a radio interferometer, between the sky and the correlator, is actually a highly non-linear one, with mixers, samplers, etc. It is testimony to the good design of interferometers that, in general, these non-linearities can be ignored when considering the broad system characteristics.

whereas for orthogonal-linear states, rotation is simply a standard Cartesian rotation matrix. In the radio context, rotation most commonly results from ionospheric Faraday rotation and parallactic angle rotation (e.g. rotation between the frame of the sky and the frame of the telescope for alt-az-mounted telescopes, etc).

Jones matrices will generally vary from one antenna to another (e.g. antenna gains and polarization leakage). Also, the Jones matrix coefficients can be a function of time and frequency (e.g. time-variant gains or leakages and bandpass functions). Because Jones matrices are combined multiplicatively, complicated systems are readily handled.

Jones matrices can be used to represent propagation and the signal paths up to the correlator. They are the polarimetric equivalent of antenna gains in non-polarimetric systems. In non-polarimetric systems, after the correlation process, the measured visibility is related to the true visibility by these antenna gains:

$$V'_{ij} = g_i g_j^* V_{ij} \tag{32–6}$$

Not surprisingly, there is an analogous situation with Jones matrices in a polarimetric system. To show this, we must introduce the so-called outer product (also known as the direct, tensor or Kronecker product). The outer product, $\mathbf{A} \otimes \mathbf{B}$, is defined as a new matrix in which each element a_{ij} of $\mathbf{A}$ is replaced by $a_{ij}\mathbf{B}$. A rather important (and easy to prove) property of the outer product is that

$$(\mathbf{A}_i\mathbf{B}_i) \otimes (\mathbf{A}_j\mathbf{B}_j) = (\mathbf{A}_i \otimes \mathbf{A}_j)(\mathbf{B}_i \otimes \mathbf{B}_j) \tag{32–7}$$

Now, returning to the correlator: the signal pairs input to the correlator for antennas i and j are $E'_i = \mathbf{J}_i E_i$ and $E'_j = \mathbf{J}_j E_j$. The outer product of these vectors is

$$\begin{aligned} E'_i \otimes E'^*_j &= (\mathbf{J}_i E_i) \otimes (\mathbf{J}_j E_j)^* && (32–8) \\ &= (\mathbf{J}_i \otimes \mathbf{J}^*_j)(E_i \otimes E^*_j), && (32–9) \end{aligned}$$

where

$$E_i \otimes E^*_j = \begin{pmatrix} E_{\mathrm{R},i}\, E^*_{\mathrm{R},j} \\ E_{\mathrm{R},i}\, E^*_{\mathrm{L},j} \\ E_{\mathrm{L},i}\, E^*_{\mathrm{R},j} \\ E_{\mathrm{L},i}\, E^*_{\mathrm{L},j} \end{pmatrix} \tag{32–10}$$

etc. With integration, the above is

$$< E_i \otimes E^*_j > = \begin{pmatrix} V_{\mathrm{RR},ij} \\ V_{\mathrm{RL},ij} \\ V_{\mathrm{LR},ij} \\ V_{\mathrm{LL},ij} \end{pmatrix}, \tag{32–11}$$

where the vector consisting of the four measured cross-correlations is known as the coherency vector.

So we see that the measured coherency vector, V'_{ij} is related to the true coherency vector, V_{ij}, by the outer product of the antenna Jones matrices:

$$V'_{ij} = (\mathbf{J}_i \otimes \mathbf{J}^*_j)V_{ij}, \tag{32–12}$$

i.e. the polarimetric equivalent of the "antenna-based gains" equation.

To calibrate an array, we need to determine (or at least estimate) the various Jones matrices of the system. Having determined them, we can invert their effect (it is just matrix inversion) and produce nominally perfect data. The way that the differing Jones matrices are determined (or estimated) varies considerably. For example, the parallactic angle rotation is computed. Antenna gains and leakages are usually solved for from calibrator observations or by self-calibration. Ionospheric Faraday rotation may be measured astronomically, by GPS measurements, or estimated by more indirect means.

We now address how the coherency vector is related to Stokes parameters. Let us define the "Stokes visibility vector" as

$$V_{\mathrm{S}} = \begin{pmatrix} V_{\mathrm{I}} \\ V_{\mathrm{Q}} \\ V_{\mathrm{U}} \\ V_{\mathrm{V}} \end{pmatrix}. \qquad (32\text{–}13)$$

Here the Stokes visibilities are samples of the Fourier transforms of the images of the Stokes parameters – they are complex-valued quantities. In a number of ways, a Stokes visibility vector can be thought of as an alternative coordinate system for the coherency vector, and so the relationship between the two is a coordinate transformation. The associated transform matrix is a 4 × 4 matrix, **S**, i.e.

$$V_{ij} = \mathbf{S} V_{\mathrm{S},ij}. \qquad (32\text{–}14)$$

Given this, we can relate the true Stokes visibility vector to the measured coherency vector:

$$V'_{ij} = (\mathbf{J}_i \otimes \mathbf{J}_j^*)\mathbf{S} V_{\mathrm{S},ij}. \qquad (32\text{–}15)$$

For orthogonal-circulars as the representation of the coherency vector,

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & i & 0 \\ 0 & 1 & -i & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}, \qquad (32\text{–}16)$$

whereas for orthogonal-linears, we have

$$\mathbf{S} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & i \\ 0 & 0 & 1 & -i \\ 1 & -1 & 0 & 0 \end{pmatrix}. \qquad (32\text{–}17)$$

Thus, to summarize this section, Jones matrices and the outer product produce a formalism which can readily, and compactly, describe the polarimetric response of an interferometer. Given a formalism to describe the response, it is apparent that, conceptually, the relevant matrices can be inverted, and we can produce calibrated Stokes visibilities.

3. Direction-Dependent Effects

Where effects are direction dependent, different sources within the field will experience different effects. These effects include antenna gains that vary with position on the sky (e.g. non-isoplanaticity) or distance from the pointing center (e.g. primary beam gain and off-axis polarimetric response). In these cases, the measured coherency vector cannot be described as a simple multiplication of the true Stokes visibility vector with a single response matrix. If we assume that the source is expressible as a sum of point sources, then the measured visibility coherency vector will be

$$V'_{ij} = (\mathbf{J}_{\mathrm{VIS},i} \otimes \mathbf{J}^*_{\mathrm{VIS},j}) \sum_k (\mathbf{J}_{\mathrm{SKY},i}(\rho_k) \otimes \mathbf{J}^*_{\mathrm{SKY},j}(\rho_k)) \mathbf{S}\, V_{\mathrm{S},ij,k}, \qquad (32\text{–}18)$$

where $\mathbf{J}_{\mathrm{SKY}}(\rho)$, which is a function of direction ρ, is the combination of all Jones matrices which depend on direction (e.g. all non-isoplanatic propagation and primary beam effects), and $\mathbf{J}_{\mathrm{VIS}}$ is the combination of all direction-independent effects (e.g. telescope signal path effects). The sum is over the set of point sources, and $V_{\mathrm{S},ij,k}$ is the true Stokes visibility vector for the kth source. Here we assume that all position-dependent effects precede the position-independent ones (remember matrix multiplication is not generally commutative!)

4. Mueller Matrices

In deriving the formalism, when integrating the correlations over an integration, we assumed that the relevant Jones matrices were constant in this averaging process. This leads us to an important point, which will be of interest in the next section. At radio wavelengths, our instrumentation and the propagation medium are such that generally we do not cause any depolarization of the signal. Depolarization is in many respects the polarimetric equivalent of decorrelation. There are a number of effects that cause depolarization. Perhaps an obvious one, though, is to set the integration time excessively long so that the polarimetric response is changing appreciably during the integration. Just as antenna-based gains are not necessarily a valid model when decorrelation occurs, so too the Jones-matrix formalism is not valid when depolarization occurs. This situation is far more common in the optical regime, so it is not surprising that a formalism has been developed to describe this.

It is perhaps convenient here to think of a total-power instrument (e.g. single-dish radio telescope or an optical telescope) which measures directly in the image domain. For such instruments, even if depolarization is occurring, there is an input/output relationship, which maps from the true to the measured Stokes parameters. The 4×4 matrix (which is real-valued) which gives this relationship is known as a Mueller matrix. For a total-power instrument which is describable by a Jones matrix, say $\mathbf{J}$, its corresponding Mueller matrix, $\mathbf{M}$, is readily determined:

$$\mathbf{M} = \mathbf{S}^{-1}(\mathbf{J} \otimes \mathbf{J}^*)\mathbf{S}. \qquad (32\text{–}19)$$

Indeed, for a system which does not depolarize, its Mueller matrix is always decomposable in this way. Generally, systems which depolarize cannot be decomposed (there are some exceptions). The optical community call systems that can be decomposed in this way "pure systems".

We note that, although a Mueller matrix consists of 16 real values, a Mueller matrix which is decomposable can contain only seven degrees of freedom. This is because the Jones matrix in the decomposition contains only four complex values or eight real values. However, because the intensity response is all that is important, the absolute phase of the Jones matrix in this decomposition is arbitrary, and so the number of degrees of freedom is actually just seven.

That there are "seven degrees of freedom" is quite different to a non-polarimetric case. The equivalent non-polarimetric analysis says that there is one degree of freedom (the overall gain) in a non-polarimetric system.

5. Limits to Polarimetric Calibration

Apart from notational elegance, the formalism allows us to comparatively easily analyze limits to polarimetric calibration and self-calibration. Here we will address the question of what is the best we can fundamentally do with various calibration observations. Mostly, this only gives a sketch of the reasoning – see Sault, Hamaker & Bregman (1996) for the gruesome details.

To do this analysis, consider an interferometer array, its correlator and imaging hardware and software as a single system. Given a source on the sky, this system forms polarimetric images on our computer display. It is quite reasonable to ask what the overall Mueller matrix of the system is. For simplicity, consider a point source at the phase center, and assume we deal only with a naturally-weighted dirty image (non-linear deconvolution is forbidden!). Because we have a point source at the phase center, to form our "image", we simply add all the measured visibilities and their conjugates. To get the overall Mueller matrix of the system, we add the responses of the individual baselines. So, the Mueller matrix will be:

$$\mathbf{M} = \frac{1}{N}\mathbf{S}^{-1}\left(\sum_{i,j,i\neq j}(\mathbf{J}_i \otimes \mathbf{J}_j^*)\right)\mathbf{S}. \tag{32–20}$$

Here the sum is over all measured baselines and their conjugate baselines (hence the output is real-valued). N is the number of baselines in the sum, and $\mathbf{J}_i$ etc... are the antenna Jones matrices.

5.1. Calibration with a snapshot observation of a point source

Assume we have a single snapshot observation of a point source calibrator, and that we have no other internal instrumental calibration. How close can we get to fully describing the polarimetric response of the array? If we assume that the baseline measurements do not individually suffer depolarization/decorrelation, then we can insist that the calibration that we derive is consistent with a system without depolarization, and that after applying calibration all visibilities must have zero phase and the same amplitude. As it has no depolarization, the system must be decomposable into the form of eq. 32–19. In this case, seven degrees of freedom remain to be specified (the seven degrees of freedom of a pure system). If we assume that the flux and polarimetric characteristics of the calibrator are unknown, the snapshot observation does not allow us to determine any of these seven degrees of freedom. These seven undetermined parameters correspond to the absolute flux scale, and to six real parameters causing leakage between the different Stokes parameter (I with Q, I with U, I with V, Q with U, Q with V and U with V). This should be compared with the non-polarimetric case, where there is only one undetermined parameter – the absolute flux scale.

Continuing in this theme, assume that we know the flux and polarization of the point source calibrator. Then, at best, we can determine four of the seven degrees of freedom (one degree of freedom for each Stokes value given), but at least three degrees of freedom must remain undetermined.

Note that this is a very general result. We have not specified the feed type (circular, linear, the so-called crossed-linear configuration possible at Westerbork, or some bizarre mix), and it does not matter whether it is an unpolarized or strongly polarized calibrator. Also note that this argument just gives the best case: we can do no better from a snapshot observation of a known calibrator than have three degrees of freedom undetermined. However, this best case is achieved in practice for arrays with at least 4 antennas and when all baselines are measured (actually the best can usually be achieved with somewhat weaker conditions). For an observation of an unpolarized calibrator, these undetermined degrees of freedom correspond to leakage between the polarized quantities (Q with U, Q with V and U with V), but not between I and the polarized quantities. This is a very reasonable result, because, clearly, an observation with an unpolarized calibrator has not probed leakage between the polarized quantities. Again for an unpolarized calibrator, but specifying an array where all the feeds are the same, these three degrees of freedom correspond instrumentally to a complex-valued offset to the D-terms and the phase difference between the orthogonal polarization channels of the array (the so-called XY or RL phase difference offset).

These undetermined degrees of freedom should be taken with some seriousness. At least for equatorially-mounted antennas, or a snapshot observation (we consider a long synthesis with alt-az antennas later), there will be no errors apparent in the resultant images (although they might be astrophysically implausible). *Polarimetrically they will just be wrong.* However, these undetermined degrees of freedom should not be taken too seriously either. Generally we can make reasonable assumptions which allow us to overcome them. For example, we could assume that the engineering of the antennas is good and that the D-terms average out to zero. This is usually quite a good approximation. Also, the array could contain hardware to measure a nominal value for the XY/RL phase difference offset (the VLA does not have such hardware, although other interferometers such as the ATCA do). Finally, because the polarized quantities are usually much smaller than total intensity, leakage between the polarized quantities is usually not as significant as leakage from total intensity.

5.2. Calibration with a long synthesis of a point source

Continuing to assume a calibrator with known polarization, we find the same argument about three undetermined degrees of freedom applies to a long synthesis with either

- an array with equatorially-mounted antennas and a calibrator with any polarization, or
- any array and a calibrator with no linear polarization.

In these cases, a long synthesis does not provide any extra information - it simply re-measures the same information, and so does not allow better polarimetric calibration (ignoring signal-to-noise issues). However, provided we assume that the instrumental parameters associated with these degrees of freedom are constant with time, a long synthesis (or many snapshots) with a calibrator of known polarization and containing some linearly polarized emission, does provide extra information for an array with alt-az mounts (or any mount where there is parallactic angle rotation). The rotation between the frame of the antenna and the frame of the sky allows the polarimetric calibration to be completely determined. Parallactic angle rotation can be thought of as providing a different calibrator. Conway & Kronberg (1969) were probably the first to fully exploit this potential in polarimetric calibration.

Instead of just allowing the telescope mount to cause rotation between the sky and the antenna, the feeds could be attached to a precision rotatable platform. Although this is still a common design in single dishes (for many reasons other than calibration), it does not appear to have been used in interferometer arrays.

5.3. Requirements for complete calibration

Assuming point sources with known fluxes and polarization, how many calibrators are needed to fully determine polarimetric calibration? Surprisingly three calibrators are needed (not two, as might be expected from simple 'counting the degrees of freedom' arguments): at least two of these must be polarized, one must be linearly polarized, and the Stokes vectors (the vector of I, Q, U and V) of the calibrators must be linearly independent. Alternatively, when the parallactic angle rotation trick can be used, observations with at least three parallactic angles are needed (in practice, many more parallactic angles than this are usually measured).

5.4. A practical case

As a final example of possible strategies, one may ask how well we can polarimetrically calibrate when we have a linearly polarized calibrator where its flux density and polarization are unknown. In practice, this is a very important case, as calibrators will often have variable flux and polarization, and so these cannot be assumed. It is also a very important case, as this is equivalent to asking what is the best we can do with self-calibration. If the parallactic angle rotation trick cannot be played, then we are in a real bind – we cannot determine any of the seven possible degrees of freedom. Thus, polarimetric calibration of equatorially-mounted antennas does force one into having faith in knowing the polarization state of the calibrators. However, if the parallactic angle rotation trick can be played, four of the seven degrees of freedom can be determined. The remaining degrees of freedom that cannot be determined correspond to the absolute flux scale, the absolute alignment of linear polarization angle and the term which causes leakage between I and V. A primary calibrator (or calibrators), with known total intensity, circular polarization and angle of linear polarization, is needed to determine these.

6. Implementation in AIPS++

We have seen how the HBS measurement equation provides a straightforward, compact theoretical description of radio-interferometric polarimetry. It can also be used for calibration and imaging of radio-interferometric polarimetry data. With this in mind, the HBS formalism has been used as the basis for synthesis processing in the AIPS++ software package now under development (AIPS++ 1998). A formalism for calibration and imaging using the HBS measurement equation has been developed and implemented in AIPS++ (see Noordam 1995, 1996; Cornwell 1995a, 1995b; Cornwell & Wieringa 1996a, 1996b; Cornwell 1996). Here we concentrate upon the formalism; for those interested, the implementation in terms of C++ classes has been described by Cornwell & Wieringa (1997).

As we have seen above, for a source expressible as a sum of point sources, the measured coherency vector is given by:

$$V'_{ij} = (\mathbf{J}_{\mathrm{VIS},i} \otimes \mathbf{J}^*_{\mathrm{VIS},j}) \sum_k (\mathbf{J}_{\mathrm{SKY},i}(\rho_k) \otimes \mathbf{J}^*_{\mathrm{SKY},j}(\rho_k))\mathbf{S}\, V_{\mathrm{S},ij,k}. \qquad (32\text{–}21)$$

The true visibility vector $V_{\mathrm{S},ij,k}$ for the kth point source is given by:

$$V_{\mathrm{S},ij,k} = I_{\mathrm{S},k} \exp(i\Phi_{ij,k}), \qquad (32\text{–}22)$$

where $I_{\mathrm{S},k}$ is the Stokes vector of the point source (i.e. the vector of the point source I, Q, U and V), and $\Phi_{ij,k}$ is the usual Fourier phase term.

The goal of calibration is to determine and correct for the Jones matrices, either from calibration observations or some other source of knowledge (e.g. GPS measurements of ionospheric total electron content, analytic models of the primary power response of the antenna, etc.). Similarly, the goal of imaging is to determine the sky brightness represented in this equation by the true Stokes vector $I_{\mathrm{S},k}$ for all positions spanning the region of interest in the image plane. Thus, one may think of one or both of the Jones matrices $\mathbf{J}$ and the Stokes vectors $I_{\mathrm{S},k}$ as unknowns, and the coherency vectors as measured (with some e.g. normally-distributed error term).

The cases of interest are:

1. $\mathbf{J}$ unknown, I_{S} known: calibration from observations of one or more calibrators of known properties.

2. $\mathbf{J}$ known, I_{S} unknown: imaging from *a priori* calibrated data.

3. $\mathbf{J}$ unknown, I_{S} unknown: self-calibration of calibration and source properties.

In practice, one often goes through all of these cases in sequence. Data are calibrated using calibrators of approximately known properties, an image is constructed from the resultant calibrated data, and then self-calibration is used to refine both the calibration and the imaging.

If we ignore the presence of the term $\mathbf{J}_{\mathrm{SKY}}$, then these steps are roughly as covered in previous lectures but with the added complication that matrices and vectors instead of scalars must be estimated:

1. Calibration: the Jones matrices $\mathbf{J}_{\mathrm{VIS}}$ can be estimated via (non-linear) least squares techniques (Cornwell 1996). The Jones matrices from calibrator observations may be interpolated in time (or other parameters) to give estimates of values on-source. The calibrated coherency vector is then estimated by inverting the product of the relevant Jones matrices:

   $$V_{ij}^{\mathrm{C}} = (\mathbf{J}_{\mathrm{VIS},i} \otimes \mathbf{J}_{\mathrm{VIS},j}^{*})^{-1} V_{ij}'. \qquad (32\text{–}23)$$

   The role of the inverse is to decouple measurements that were coupled by some instrumental effect, such as for example, parallactic angle rotation for an alt-azimuth mounted antenna, or polarization leakage.

2. Imaging: In previous lectures, it was shown that estimating the point source brightnesses via a least squares method leads to an under-determined linear equation involving a dirty image and a dirty beam. A similar argument in this case leads to a similar convolution equation in which each pixel of the dirty image is a vector of I, Q, U and V:

   $$I_k^{\mathrm{D}} = \mathrm{Re}\left[S^{*T} \sum_{ij} w_{ij} V_{ij}^{\mathrm{C}} \exp(-i\Phi_{ij,k}) \right], \qquad (32\text{–}24)$$

   where the w_{ij} are appropriately normalized weight terms. The brightnesses of the point sources can be estimated by one of a number of deconvolution algorithms: CLEAN, Maximum Entropy, etc. The deconvolution algorithms can proceed independently for each of I, Q, U, V or the deconvolution may be coupled in polarization by, for example, in CLEAN, searching for peaks in $\sqrt{I^2 + Q^2 + U^2 + V^2}$ rather than I alone (see Cornwell 1995 for further discussion).

3. Self-calibration: If both the calibration and the sky brightness are unknown then one has to solve a complicated set of non-linear, under-determined equations. Although this sounds hard, a straightforward iterative approach works nearly all the time. The calibration and imaging steps are iterated, solving for calibration and the image alternately, while applying image plane constraints, such as $I \geq \sqrt{Q^2 + U^2 + V^2}$, to limit the degrees of freedom. A subtle point is that since the calibration effects are factorized per antenna via the Jones matrices, an analog of the closure relations is automatically obeyed. Thus one can think of the HBS measurement equation as allowing generalized self-calibration.

So, in the case of no direction-dependent effects, calibration, imaging, and self-calibration can be straightforwardly derived for the HBS measurement equation. The resulting algorithms are conceptually and structurally similar to those for the conventional, non-polarimetric formulation of interferometry. However, if direction-dependent effects must be taken into account, then the notions of calibration and imaging become intertwined. Physically, this makes sense, because if, for example, an antenna-based gain or phase term varies as a function of direction then this inevitably gets confused with the signature of source structure. Mathematically, the simple step of correcting the calibration by pre-multiplying by the inverse of the direct product of Jones matrices no longer works.

To investigate further the case of direction-dependent effects, let us first turn to imaging in the presence of *a priori* known image-plane calibration terms. A simple example would be a model for the primary beam of the array antennas. Generally, when direction-dependent effects are present, the resultant point-spread function will be shift-variant, and so traditional deconvolution approaches cannot be used. Cornwell (1995) shows that one viable approach is to use an algorithm in which an estimate of the sky brightness is iteratively improved. A plausible update formula can be derived from the gradients of the fit (χ^2) of the predicted coherency vectors to those measured. Using an approximate Newton-Raphson approach, in which the diagonal elements of the second derivative of χ^2 with respect to $I_{\mathrm{S},k}$ are taken into account, one can define a generalized residual image:

$$I^D_{\mathrm{S},k} = -\left[\frac{\partial^2\chi^2}{\partial I_{\mathrm{S},k}\partial I^T_{\mathrm{S},k}}\right]^{-1}\frac{\partial\chi^2}{\partial I_{\mathrm{S},k}} \qquad (32\text{–}25)$$

where

$$\frac{\partial\chi^2}{\partial I_{\mathrm{S},k}} = -2\,\mathrm{Re}\left[\sum_{ij}\left[\mathbf{J}_{\mathrm{SKY},i}(\rho_k)\otimes\mathbf{J}^*_{\mathrm{SKY},j}(\rho_k)\right]^{*T}\Lambda_{ij}\;\Delta V_{ij}\exp(-i\Phi_{ij,k})\right], \qquad (32\text{–}26)$$

$$\frac{\partial^2\chi^2}{\partial I_{\mathrm{S},k}\partial I^T_{\mathrm{S},k}} = 2\,\mathrm{Re}[H], \qquad (32\text{–}27)$$

$$H = \sum_{ij} S^{*T}\left[\mathbf{J}_{\mathrm{SKY},i}(\rho_k)\otimes\mathbf{J}^*_{\mathrm{SKY},j}(\rho_k)\right]^{*T}\Lambda_{ij}\left[\mathbf{J}_{\mathrm{SKY},i}(\rho_k)\otimes\mathbf{J}^*_{\mathrm{SKY},j}(\rho_k)\right]\;S\,, \qquad (32\text{–}28)$$

and Λ_{ij} is the inverse of the covariance matrix of measurement errors on the coherency vector, and ΔV_{ij} is the residual coherency vector (i.e. prediction error of the coherency vector).

Thus, the residual image is approximately a Fourier summation of the residual coherency vectors, modified by multiplication with the adjoint of the image-plane effect. The residual is normalized by an appropriate sum of the self-products of the image-plane effects. Although this updated formula looks quite fearsome, it is actually quite straightforward and does reduce to known cases. For example, if the $\mathbf{J}_{\mathrm{SKY}}$ terms represent (only) the electric field reception pattern of the antennas, then the normal mosaicing equations (see e.g. Cornwell, Holdaway & Uson 1993) are obtained. Hence, it would be reasonable to think of this as a generalization of mosaicing in which not only the antenna primary beams can be corrected but any factorizable image plane effect can be corrected.

Note that since this is just an update formula, it must be used in an iterative algorithm to estimate the sky brightnesses $I_{\mathrm{S},k}$. In general, this formula cannot be reduced to a shift-invariant convolution equation, although an approximate convolution relation may be feasible. If so, then an algorithm like the following can be used:

1 Estimate starting model image $I_{\mathrm{S},k}$

2 Calculate residual image for the current model image using above formula, an approximate point spread function, and a threshold in brightness below which the approximation is no longer valid.

3 Clean the residual image using the approximate point spread function down to the threshold.

4 Return to step 2 and continue if the remaining peak residual is too high.

5 Smooth the set of point sources thus found and add the residual image to form a restored image.

If the Jones matrices $\mathbf{J}_{\mathrm{SKY}}$ change sufficiently infrequently then FFT-based convolutions can be used to speed the processing required in the calculation of the residual image. In this case, the algorithm is very similar in structure to that proposed for mosaicing by Sault, Staveley-Smith & Brouw (1996).

If the direction-dependent terms $\mathbf{J}_{\mathrm{SKY}}$ are not known *a priori* then a gradient search approach may be used to estimate them. We have little practical experience with this approach but general arguments would indicate that this would be very difficult to realize unless a sufficiently simple parameterization of the Jones matrices is used.

The actual implementation of these mechanisms in AIPS++ has a number of refinements that add power and reduce computing costs. Description of these may be found in the paper by Cornwell & Wieringa (1997), and in the AIPS++ On-line Documentation (AIPS++ Project 1998).

References

AIPS++ Project 1998, http://aips2.nrao.edu/aips++/docs/html/.

Conway, R. G., & Kronberg, P. P. 1969, *MNRAS*, 142, 11–32.

Cornwell, T. J., Holdaway, M. H., & Uson, J. M. 1993, *A&A*271, 697–713.

Cornwell, T. J. 1995, *The Generic Instrument: I Overview of Calibration and Imaging*, AIPS++ Note 183, http://aips2.nrao.edu/aips++/docs/notes/183/183.html.

Cornwell, T. J. 1995, *The Generic Instrument: II Image solvers*, AIPS++ Note 184, http://aips2.nrao.edu/aips++/docs/notes/184/184.html.

Cornwell, T. J. & Wieringa M. H. 1996a, *The Generic Instrument: III Design of Calibration and Imaging*, AIPS++ Note 189, http://aips2.nrao.edu/aips++/docs/notes/189/189.html.

Cornwell, T. J. 1996, *The Generic Instrument: IV Specifications and Development Plan*, AIPS++ Note 192, http://aips2.nrao.edu/aips++/docs/notes/192/192.html.

Cornwell, T. J., & Wieringa M. H. 1996b, *The Generic Instrument: V Design of Cross-calibration*, AIPS++ Note 193,
`http://aips2.nrao.edu/aips++/docs/notes/193/193.html`.

Cornwell, T. J., & Wieringa, M. H. 1997, in *Astronomical Data Analysis Software and Systems IV* eds. G. Hunt & H.E. Payne (San Francisco: PASP), 10–17.

Hamaker, J. P., Bregman, J. D., & Sault, R. J. 1996, *A&AS*, 117, 137–147.

Hecht, E., & Zajac, A. 1974, *Optics*, Addison-Wesley, Reading, MA.

Mayer, C. H., McCullough T. P., & Sloanaker R. M. 1957, *ApJ*, 126, 468–470.

Noordam, J. E. 1995, *Some practical aspects of the matrix-based Measurement Equation of a generic radio telescope*, AIPS++ Note 182,
`http://aips2.nrao.edu/aips++/docs/notes/182/182.html`.

Noordam, J. E. 1996, *The Measurement Equation of a Generic Radio Telescope*, AIPS++ Note 185,
`http://aips2.nrao.edu/aips++/docs/notes/185/185.html`.

Pawsey J. L., & Bracewell R. N. 1955, *Radio Astronomy*, Oxford Univ. Press, London, England.

Sault, R. J., Hamaker, J. P., & Bregman, J. D. 1996, *A&AS*, 117, 149–159.

Sault, R. J., Staveley-Smith, L., & Brouw, W. N. 1996, *A&AS*, 120, 374–384.

Tinbergen, J. 1996, *Astronomical polarimetry*, Cambridge Univ. Press, Cambridge, UK.

Turlo, Z., Forkert, T., Sieber, W., & Wilson, W. 1985, *A&A*, 142, 181–188.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

33. Noise and Interferometry

V. Radhakrishnan
Raman Research Institute, Bangalore; and Institute of Astronomy, Amsterdam

Abstract.
The nature of noise and its varied manifestations in radio and optical interferometry are the subject of this lecture.

1. Preamble

All of the lectures that you have had until now dealt with the theory and practice of how to make images of the radio radiation from the sky, and by now you are probably experts at it. A simplified statement of the principle used is that the similarity between the signals received at spatially separated points contains the information which suitably transformed renders an image of the sky. Although many telescopes are generally used simultaneously in such an exercise, it is the pairwise comparisons that contain the information and that are then put together. It is the similarity between the signals received in two antennas as a function of their vector spacing that is quantified by measuring their complex correlation coefficient. This coefficient will vary from one pair of antennas to the next and measurements on many such pairs go into making the final image. But one pair of telescopes or antennas will suffice for this lecture, whose main purpose is to look a little closer into the meaning of the similarity of signals.

In this last lecture of the school, I shall touch upon a number of topics which have no immediate or everyday relevance to the use of radio interferometers for synthesis imaging, but an appreciation of which could provide greater insight into what your have learnt so far. Towards this end, the following list of items was suggested to me by the organizers of the school:

- The nature of noise.
- Weak and strong signals.
- The two slit paradox.
- Classical versus quantum regimes.
- Relation to intensity interferometry.
- Detection thresholds.
- Amplifiers and amplifier noise.
- Optical versus radio concepts.
- Photon statistics in different regimes.

My talk will attempt to cover all of them, but not necessarily in this sequence.

2. Information and Bandwidth

Noise is unfortunately a bad word for a very good thing like the radiation that emanates from celestial sources bringing us the information we seek. There are also unwanted kinds of noise, like receiver noise that is generated within the receiver itself, or noise from the ground due to spill-over into the antennas. So when I say noise, I just mean a signal with the properties of natural radiation like that which comes to us from radio sources in the sky, or the thermal radiation from a resistor. It is very different for example, from the radiation that you get from a signal generator in the laboratory, whose energy is delivered in such a narrow frequency interval that we can think of it as having no width at all. The latter is very similar to the electric mains where all of the power is in an extremely narrow band (around 60 Hz in this country).

If you take such a source of monochromatic radiation, as it is called, its power and frequency are both fixed for ever. We can learn nothing from it as a function of time, because nothing whatever changes. It is true that the phase of the signal goes round and round, but it does so in a totally predictable fashion and one needs to measure the phase only once to know it precisely at any given future instant. This type of radiation carries no information and as an illustration, let us imagine that a particular radio source in the sky puts out such a signal. If you applied all of the techniques that you have learnt in this synthesis imaging school, all you could obtain, apart from its intensity and polarization, would be its coordinates.

On the other hand, the information needed to make all the lovely images of different radio sources that you have seen is carried in the bandwidth associated with their radiation. It is such broadband signals, as opposed to monochromatic radiation, that we call Noise and that has wonderful properties. Although there are so-called "line" sources that radiate spectral lines like the interstellar hydrogen or OH, as distinct from "continuum" sources, they are all the same for our purpose. If you look at the radiation in a band that is small compared to the actual width of the spectral line, then the characteristics of the signal would be the same as those in a continuum source observed at the same frequency and within the same bandwidth. So what are the characteristics of natural radiation?

3. The Nature of Noise

There are two ways of looking at such a signal which are really equivalent. One is to look at the signal in time and the other in frequency space. If you look at it in time (say on an oscilloscope) you will find that it varies from instant to instant in a way that its average power gets closer and closer to some value the longer the integration time. But the exact value at some future instant remains unpredictable. A signal with a small fractional bandwidth will look like a regular sinewave of approximately the frequency of the center of the band, and in a time of the order of the reciprocal of the bandwidth, its frequency and amplitude change to new values. If the bandwidth is large then these changes happen quickly. If the bandwidth is small then the changes happen more slowly. The data rate is given by the bandwidth, each new piece arriving in a time that is its reciprocal. This is also the rate at which the state of polarization of natural radiation can change to a new state. And finally, an important characteristic of noise is that its distribution in amplitude is Gaussian.

Turning to frequency space, how do we know that the signal is truly broad band? Let us begin by just splitting the band into two halves. We will now find that the average powers are equal, but that the instantaneous value of the signal in the left hand half of the band is quite independent of that in the right hand half, with no correlation whatsoever. We also find that the signals change half as fast as before because the bandwidth has been halved. This suggests that there is signal power at all frequencies and to prove this we could go on subdividing the bands only to find that no matter how narrow we make them, there is always mean power there proportional to the bandwidth.

As noted already we also find that both the amplitude and the phase change more and more slowly as the bands are made narrower and narrower, but the amplitude distribution remains Gaussian however narrow the bands. This leads us to another way of looking at the broad band signal, namely that it consists of an infinite number of monochromatic signals occupying every possible position within this band. Each one of these monochromatic waves is, of course, totally predictable and cannot change for reasons we have already discussed earlier. But when the passband allows a certain range of them to come through, they are all going at different rates and the value of the voltage that we measure at some instant simply happens to be that given by the instantaneous addition of all these infinite monochromatic components. Thanks to the central limit theorem this ensures a Gaussian distribution. At the next instant the voltage will be slightly different and it will go on changing. But the most rapid change possible is that due to the rates associated with the extreme frequencies in the band, and that is why independent amplitudes and phases are separated by time intervals of the order of $(1/\Delta\nu)$. And also why it is enough to sample a band-limited waveform at a finite rate to reconstruct it completely.

4. Interferometers & Coherence

Having described noise, it is time to turn to interferometers again and to see what noise does to them. If a radio source, like the hypothetical one, that I mentioned a little earlier, radiated only a monochromatic wave then the correlation coefficient that would be measured by a pair of antennas receiving the signal would clearly be 100%. We could also say that the signals received by the two antennas were totally coherent.

Now if we did this experiment with real telescopes, for example a pair of antennas of the Very Large Array, the rotation of the earth during our measurement would Doppler shift the apparent frequencies received at the two antennas, offsetting one from the other. This would surprise nobody and the matter would simply be described as saying that we have a fringe rate, meaning thereby that the monochromatic signals received at the two telescopes differed by this amount in frequency. It is this frequency difference and its variation with time which allowed us to measure the source position. One would however still continue to find that the magnitude of the correlation coefficient was 100%, i.e., that the signals received at the two telescopes were fully coherent. This is a simple and convincing way of appreciating that any monochromatic signal is totally coherent with any other monochromatic signal whatever their frequency difference. And it is also a very good way to see that incoherence between two signals requires them to have a finite bandwidth. But broadband signals can also be fully coherent as we shall soon discuss.

Going back to noise, if we look within the time represented by a reciprocal of the bandwidth, then, as I said earlier, each of the antennas will be receiving something that is temporarily monochromatic but which will differ in frequency and amplitude in the two antennas. Nonetheless, for reasons just given, the correlation coefficient for these two should again be unity, and these two signals have to be considered coherent during this small interval. The notion of partial coherence or of a correlation coefficient that is less than 100% can arise only when we combine a large number of samples, each of which can have different values for the phase, finally converging to an average which could be anything between 100% and something close to, but never identically zero.

There is a deep analogy between this and the measurement of the degree of polarization of natural radiation, not least because polarization is a measure of the correlation between orthogonal components of the electric (or magnetic) fields. Any individual sample we measure will be found to be 100% polarized with some particular polarization state even when observing a black body. When we average a large number of samples, and find that the state of polarization of the different samples is distributed uniformly in polarization space, we say that the radiation is randomly polarized. On the other hand, if there is a bias towards any particular state, the radiation is said to be partially polarized. The degree of polarization like the degree of coherence is a notion that can be entertained only when we talk about broad band signals. Thus we see that NOISE is the very stuff of interferometry. Its characteristics provide all the consequences of measuring correlations of the radiation field sampled at locations separated in space and/or time.

5. The Similarity of Signals

We can now look a little closer at examples of cases where there is little, or much, similarity between two signals. If we think of the noise from two different resistors, it is easy to persuade ourselves that their signals will be totally uncorrelated because the microscopic processes that agitate the electrons in these resistors have nothing to do with each other. In the same way, the radiation received from two separated radio sources in the sky or generated in two different receivers will be totally independent and uncorrelated. So also the radiation from separated pieces of ground that different telescopes might see.

A totally different example would be to take a signal that is propagating along a pair of wires and to split it into two. No one would doubt that if you took the output from one of the audio channels of a music system and split it to drive two loud speakers that you would hear the same music from both. Or, that if you split the signal from one resistor into two and put them into a correlator, that you would get an answer of 100% . There is also no difficulty in appreciating why signals which have been derived from a common source, but have different incoherent additions to the two halves (e.g. different receiver noises), will look partially dissimilar. One of today's exercises however is to see how without the addition of extra uncorrelated noise, the division of signals coming from a single source can give unlike samples.

We know that the signal received from a radio source by a telescope has large fluctuations in phase and amplitude because, as already discussed, the signal is broadband or NOISY. Also that it is this noisiness which gives meaning to any measurement of similarity between two signals, and that new information comes only in samples whose rate is given by the bandwidth. But what is it that governs the similarity of the samples received by two telescopes from one source?

Now this is the topic on which, as I said, you must all be experts by now. What you have learnt is that even though a distant radio source might subtend a very small angle in the sky, it is nevertheless composed of different regions, each of which radiates independently. And therefore, when you have separated receiving points, the ways in which the contributions from the different regions combine are not the same at the two telescopes and this results in a correlation coefficient of less than unity.

We can now do a little thought experiment and somehow move the source further away, for example by imagining the expansion of the universe to have gone a lot faster than it does. What we want is the source to retain its characteristics but to simply get more and more distant from the interferometer. Let us see what this does to the (degree of) similarity of the signals received at the two antennas. What is certain is that the angular diameter of the source will decrease as it gets further and further away from us, and it will therefore be less and less resolved by the interferometer. As a consequence, the degree of correlation would definitely increase if this were the only effect. But although it appears more and more like a point source, distance will eventually make it so weak that the division of its signal between the two telescopes will introduce dissimilarity between the two versions. Granularity will appear simply because radiation cannot be divided into pieces smaller than photons. When will this happen?

6. Wave Noise

One way that I like to visualize this is with a long train of boxcars, each piled up with some number of objects which could be bricks or atoms, or in this case photons. At a certain point, the contents of each boxcar has to be divided between two others belonging to two different trains whose pattern of contents we are then going to compare. Let the original train have some enormous number of bricks in each boxcar as in Figure 33–1A. When the contents of each boxcar are divided between two let us assume that the bricks are not counted out but individually thrown randomly into the boxcars of the two trains. The difference between the piles in any corresponding pair of boxcars will be typically of the order of the square root of the number of bricks in each of them; and therefore, the larger the original number, the more exactly similar would be the two portions into which it is divided. Consequently, the pattern of variation of the number of bricks along the trains would be very similar and give us almost perfect correlation of their sizes.

This corresponds to the case of splitting an audio signal into two loudspeakers, and represents the classical limit. What characterises such a classical signal is an enormously high photon occupation number in each sample which permits it to be split into two or more almost identical versions. Even more important is that in the classical limit the voltage (or current) waveform can be measured as a function of time without the measurement process introducing significant error. This implies the ability to record and reproduce a signal with fidelity and is what permits post facto comparison of signals for similarity. VLBI would not be possible without this ability. Its implications, and the price that has to be

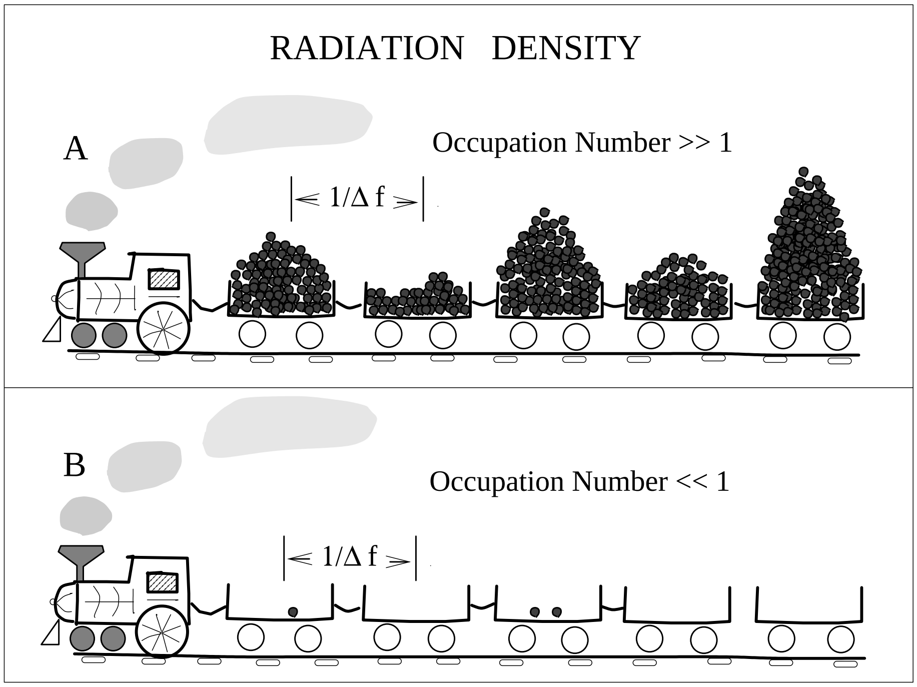


Figure 33–1. Boxcar representation for a stream of radiation. Each boxcar is a sample and corresponds to the reciprocal of the bandwidth, the rate at which new information arrives. A) The high density case where there is an enormous number of photons in each sample and substantial variation from sample to sample. B) The very low density case when the number of photons is minute compared to the number of samples.

paid for it, will become clear when I discuss amplifiers towards the end of my talk.

To avoid confusion with amplitudes, let me point out that the variations from one pile of bricks (photons) to the next as seen in Figure 33–1A represent intensity changes with a time scale of the inverse of the bandwidth. These variations are called wave noise and are associated with high density signals. If the bandwidth is say a hundredth of the frequency, then there are of the order of a hundred cycles of the waveform within one sample. In conventional interferometry, it is the amplitude and phase of this waveform that are correlated with those from other telescopes as you have learnt. The correlation of just the size of one pile with that of another is the correlation of intensities which I shall discuss shortly.

7. Shot Noise

As the other extreme, you could imagine a train in which the number of bricks is minute compared to the number of boxcars, Figure 33–1B. This means that most of them are empty and that occasionally you find a boxcar with a single brick, and even less often a boxcar with more than one. This situation is very different to the picture I have painted above. Firstly, the rate at which the signal brings us new information is not the bandwidth, but rather the mean number of bricks (photons) per second received by the telescope. Then again, if such a sequence consisting mostly of zeros were split into two as we did before, we must surely end up with all the ones leading to a one and a zero in the daughter sequences. This looks like perfect anticorrelation, and suggests that interferometry as I have described it so far, cannot work with such weak signals.

Before discussing how it can and does, let me assure you that although it may seem strange to radio astronomers, this is a faithful description of signals from astronomical sources at optical and all higher frequencies. A square meter of collecting area looking at a zero magnitude star receives on average one optical photon for every ten thousand sample times. The separations get even greater as we go to higher frequencies like X-rays. The arrival of photons at random times with a mean separation that is orders of magnitude greater than the inverse of the filter bandwidth gives rise to the very different, and shot-like noise associated with all such sparse signals. The quantization of the received energy is evident, and it is the associated photon noise, characterized by Poisson statistics, that is responsible for the uncertainty in measurements of such weak intensities.

A telling illustration of the difference between the two cases is a comparison of the accuracy of intensity measurements of radio and optical signals as a function of telescope size. When the size of a radio telescope is big enough to give an antenna temperature on a source that is the main contribution to the total noise, further increase of telescope size makes no improvement even though the source may be totally unresolved by its beam. The uncertainty in a measurement of the intensity depends only on the bandwidth and the integration time, whose product gives you the number of samples, and the square root of which determines the fractional error, Figure 33–2A. In the case of an optical telescope measuring the intensity of light from a star, also unresolved, the square root is of the number of photons detected, which depends on the collecting area and continues to increase with telescope size! (Figure 33–2B).

8. Intensity Interferometry

The possibility of correlating intensities as mentioned a little earlier has played an important part in the development of astronomical interferometry, and also in the launching of a new branch of physics called quantum optics. Two of the most spectacular synthesis images of radio sources that you have surely seen are of Cas A and Cygnus A made here by some of your teachers in this school. The first attempts to resolve these sources at radio wavelengths was in the early fifties by Hanbury Brown and colleagues at Jodrell Bank. It was then believed that these sources might be so small that baselines as large as are now used in the VLBA might be needed to resolve them, and so Hanbury Brown invented, or discovered, a way to do this without a coherent local oscillator at both receivers. It was simply to compare the fluctuations in the demodulated waves which was done by transmitting the detected noise on a radio link. The price paid is a lower signal to noise ratio, which is a small sacrifice if it enables an otherwise impossible observation. The scheme worked perfectly, but to their disappointment it all ended too soon as both sources were resolved with baselines

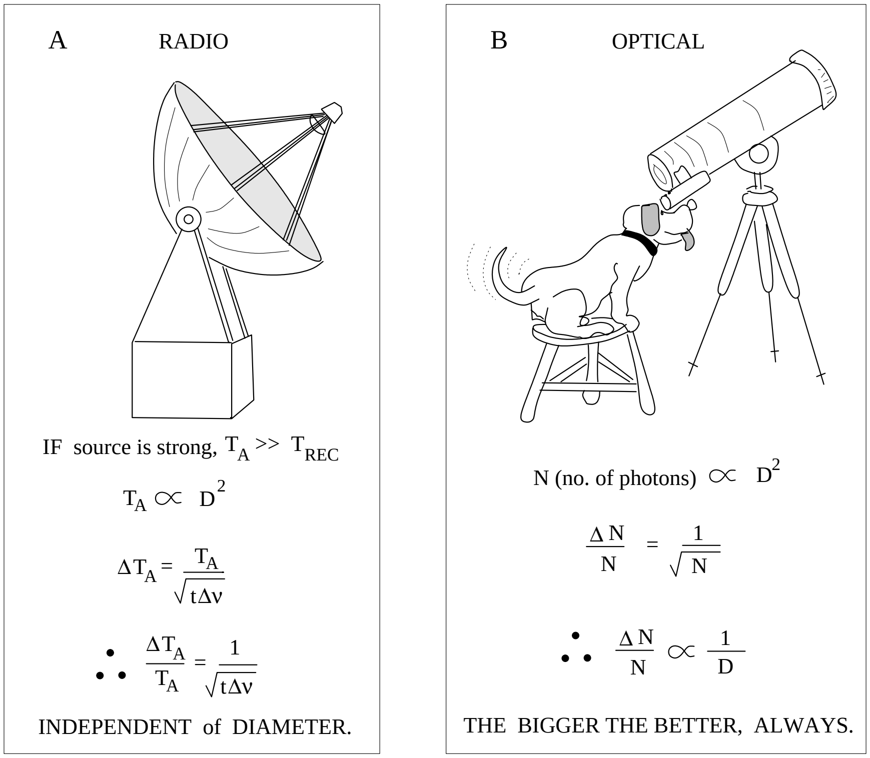

A
RADIO
IF source is strong, $T_A >> T_{REC}$
$T_A \propto D^2$
$\Delta T_A = \frac{T_A}{\sqrt{t\Delta\nu}}$
$\therefore \frac{\Delta T_A}{T_A} = \frac{1}{\sqrt{t\Delta\nu}}$
INDEPENDENT of DIAMETER.
B
OPTICAL
N (no. of photons) $\propto D^2$
$\frac{\Delta N}{N} = \frac{1}{\sqrt{N}}$
$\therefore \frac{\Delta N}{N} \propto \frac{1}{D}$
THE BIGGER THE BETTER, ALWAYS.

Figure 33–2. The surprising difference in the dependence on telescope size of the accuracy of intensity measurements of radio and optical signals, corresponding typically to the high and low density cases illustrated in Figure 33–1. In the case of a strong radio source, A), the fractional error depends only on the square root of the number of samples and is independent of telescope size. B) In the optical case, where the stellar radiation is the only input, it is the number of photons which matters, as null samples carry no information; the accuracy therefore continues to increase with telescope size. In the general case, the quantity whose square root determines the error in the intensity measurement is the harmonic mean of the number of photons and the number of samples.

of only a few kilometers, and did not really need a scheme which could work across the world.

In Hanbury's words they had used a sledge-hammer to crack a nut, but luckily there were more rewards to come. The fact that the correlations were not affected even when the radio sources were scintillating violently due to the ionosphere revealed that the system could also be made to work through a turbulent medium. And so was born the idea of making an optical intensity interferometer with baselines long enough to resolve main sequence stars. Michelson, the pioneer in this field could do no better than 20 feet, the major difficulty being the turbulence in the earth's atmosphere which blurs the image into a patch which is enormous compared with the true angular size of the star. Apart from being unaffected by this turbulence, another great virtue of comparing intensities was that the mechanical stability needed was only of the order of the reciprocal of the bandwidth of the filter following the detector, about a foot. In the case of a Michelson type interferometer, the stability needed is of the order of a wavelength of light!

A measure of the originality of any idea in science is often the opposition it evokes, and Hanbury Brown and his theoretician-collaborator Twiss got more than their fair share. But they gave as good as they got, and finally managed to obtain funds and to build an instrument in Australia with a maximum baseline of almost two hundred meters. They also achieved their principal scientific objectives. Reasonably precise and reliable measurements were made of 32 single stars; these included the first measurements ever made of a main sequence star and the first measurements of any star earlier than type M. The number of known angular diameters was increased from 6 to 38 and this work stands as a permanent and valuable contribution to stellar astronomy. In addition, the results of their observations on the spectroscopic binary Spica were a striking demonstration of the value of a high resolution interferometer for the study of a close binary star.

From my remarks in the previous section on the sparseness of optical radiation from stars, it might appear that it would be impossible to find correlations in the arrival times of photons at two separated sites. And everybody continued to say so to Hanbury Brown and Twiss till they proved otherwise. To understand why it works you should have heard the Jansky lecture that Hanbury Brown gave here some years ago, or better, read his book on the subject which is also full of his humour. Among the various reasons are that the stars chosen were hot, the reflectors large, and that photons obey Bose statistics. This tends to clump them and in a sense provides a tiny amount of wave noise which is correlated in both detectors as opposed to the shot noises which are uncorrelated.

9. Photons and Interference

I come now to a discussion of those aspects of radiation related to wave-particle duality that have provided a continuing source of confusion ever since quantum theory came into being. From ripples on a pond to Newton's rings, waves have always been the basis of understanding interference phenomena. It is the relative phase of two overlapping waves which decides whether the addition is constructive or destructive and gives rise to all of the effects you have studied at this school. If the picture of radiation as waves with amplitudes and phases is replaced by one in which the energy arrives in discrete bundles, the notion of interference becomes hard to visualize. The principal reason for this difficulty is the misleading concept of photons as particles, and miniscule ones at that. The click of a loudspeaker announcing the arrival of a photon on the tiny area of the cathode of a phototube or the even smaller area of a pixel of a CCD camera, gives an overwhelming but erroneous impression of a photon as a highly localized object in space and time.

It is this false picture of localisation that is responsible for much of the confusion in the discussion of the (in)famous two-slit paradox. Figuring out which of two holes a photon went through is very reminiscent of the preoccupation in an earlier age of assessing how many angels could stand on the head of a pin. Confusion has reigned despite the physicists who created and understood quantum theory telling us that the propagation of radiation, whatever its strength, is always wavelike. And that the discrete nature of the energy manifests itself ONLY in the process of emission or absorption, as for example when it produces a photo-electron. As a test of its correctness we shall try applying this prescription to the functioning of an optical telescope to see if it provides a consistent picture of the behavior of radiation, even if it is not an easily visualizable one.

The diameter of the reflector and the accuracy of its surface determine the collecting area and the resolution of the optical telescope just as for a radio telescope. Ignoring the effect of the atmosphere (which is not relevant to this discussion) the size and shape of the diffraction image will be just the same as would be obtained with intense classical electromagnetic radiation. But I have already mentioned that the signal received by such a telescope from a star is very weak, and consists of photons whose arrival times appear well separated. How does this manifest itself? In several ways. The first is that the exact arrival time of the next photon is totally unpredictable. The next is that the diffraction image is also built up in an unpredictable way, but always mysteriously evolving into the calculated one when enough photons have been accumulated to overcome the granularity. What should be even more surprising is that no matter how weak one makes the radiation, say by moving the source further away as we did before, the image will be the same when we have accumulated the same (adequate) number of photons. How shall we understand this?

10. Uncertainty and Probability

First of all, the notion of intensity has to be replaced by a measure of the probability of detecting a photon within a certain time interval. This applies individually and independently to different parts of the image, and weakening of the radiation would increase proportionately the average time between photons in every part of it. But the fact that the eventual accumulated diffraction pattern is the same, no matter how large this separation in time, is proof that every photon, so to speak, "senses" every part of the telescope aperture. Blocking any small area, by for instance sticking a postage stamp on the mirror, dramatically manifests itself by pushing photons into the nulls of the previous unobstructed aperture diffraction pattern.

This sensing of the whole aperture holds even if there were an opaque strip across its middle separating it in two. The corresponding diffraction pattern is of course different, and has fringes in it, but its building up photon by photon would still be independent of the intensity. It makes no more sense to ask which half the photon went through, than it would have been before to ask where exactly in the undivided aperture any photon went through. Interference can be thought of as the name given to diffraction from a non-contiguous aperture and the photon really only interferes with itself if you want to think of it that way. If you do, you must also think of its size and shape as that of the total aperture in every detail (including separations) that you allowed it to go through before its detection and consequent annihilation.

The replacement of certainty by probability when dealing with the behavior of photons is completely described by Heisenberg's celebrated Uncertainty Principle. It identifies pairs of associated quantities, an accurate determination of one of which automatically implies a large error in the other, the product

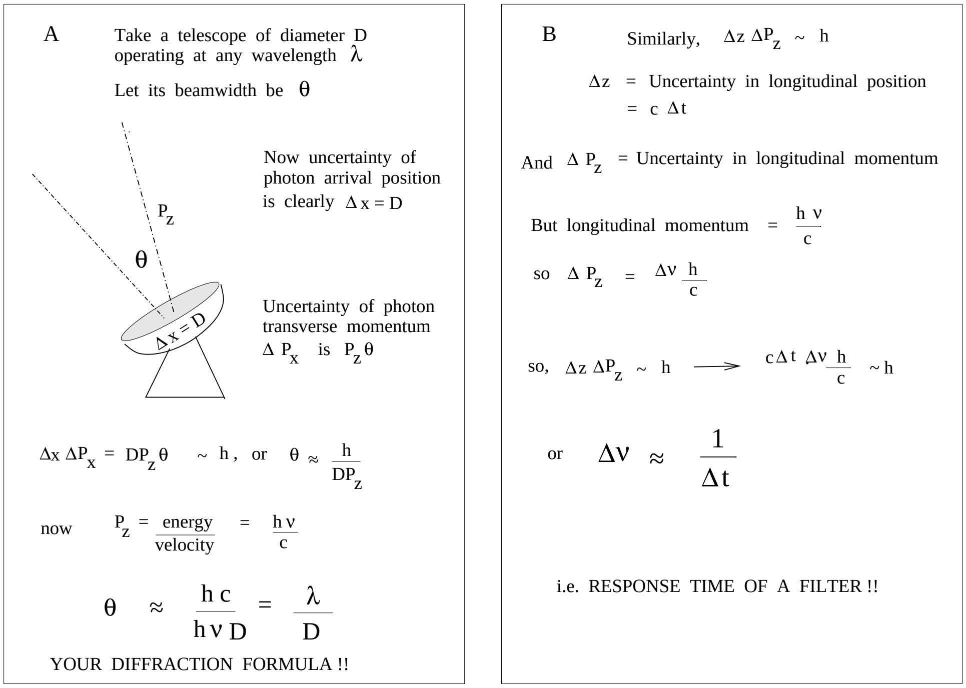

A
Take a telescope of diameter D operating at any wavelength λ
Let its beamwidth be θ
P_z
θ
$\Delta x = D$
Now uncertainty of photon arrival position is clearly $\Delta x = D$
Uncertainty of photon transverse momentum ΔP_x is $P_z \theta$
$\Delta x \, \Delta P_x = D P_z \theta \sim h$, or $\theta \approx \frac{h}{D P_z}$
now $P_z = \frac{\text{energy}}{\text{velocity}} = \frac{h\nu}{c}$
$\theta \approx \frac{hc}{h\nu D} = \frac{\lambda}{D}$
YOUR DIFFRACTION FORMULA !!
B
Similarly, $\Delta z \, \Delta P_z \sim h$
Δz = Uncertainty in longitudinal position $= c\,\Delta t$
And ΔP_z = Uncertainty in longitudinal momentum
But longitudinal momentum $= \frac{h\nu}{c}$
so $\Delta P_z = \Delta\nu \frac{h}{c}$
so, $\Delta z \, \Delta P_z \sim h \longrightarrow c\Delta t \, \Delta\nu \frac{h}{c} \sim h$
or $\Delta\nu \approx \frac{1}{\Delta t}$
i.e. RESPONSE TIME OF A FILTER !!

Figure 33–3. Heisenberg's uncertainty principle connects naturally, A) the aperture size and resolving power of telescopes, or B) the bandwidths and response times of filters, even when dealing with single photons.

of the uncertainties not being less than Planck's constant. This quantifies the informational price to be paid for localization in space or time for example, by connecting naturally the aperture size and resolving power of telescopes, or the bandwidth and response times of filters, even when dealing with single photons. The smaller the physical dimensions of a telescope the greater is the accuracy of localization in lateral position of the photon that is collected by it. The price paid, as specified by Heisenberg, is greater uncertainty in determining the lateral component of its momentum, resulting (in our language) in a wider beam, Figure 33–3A. It is the precisely analogous relation between the localization in longitudinal position and longitudinal momentum that connects bandwidth and the time resolution of signals, Figure 33–3B.

The above examples relate to the pair involving position and momentum, and another involves energy and time. A third pair which leads us back to the topic of samples is that relating phase and number (see Lecture 28). A classical signal with a well defined phase is characterized, as seen earlier, by a very large number of photons per sample, and a proportionately large uncertainty in their actual number. At the other extreme is a signal consisting of occasional photons for which the notion of an absolute phase is essentially meaningless. But relative phase differences for different possible paths continues to be meaningful and is what enables us to understand how a photon "interferes with itself."

11. Density in Phase Space

When discussing samples earlier we saw that when subdividing a band into narrower and narrower channels, the mean power available decreases in proportion to the width of the channel. A matched resistor connected to a cable is a good source of natural radiation with all the noise-like properties that I have already discussed, and we know that the power available from it is proportional to $k_B T \Delta\nu$, where k_B is Boltzmann's constant and T is the physical temperature of the resistor in Kelvins. While the power available decreases with bandwidth, we also saw that the rate at which one obtains independent samples goes inversely as the width of the band. Consequently, the energy in each sample is independent of the bandwidth and the average size of this bundle of energy depends only on the temperature characterizing the radiation. And this is equal to $k_B T$ at all frequencies which lie on the lower or Rayleigh-Jeans part of the spectrum.

As the energy of a photon can only be $h\nu$, it is trivial to calculate the number of photons in each sample. And the answer is that when the temperature of the resistor has a value of $h\nu/k_B$, the samples have of the order of one photon in each of them! At temperatures which are very high compared to this value, the signal will appear to be classical and there will be no difficulty in dividing it without introducing dissimilarity. This is the loud speaker case where I will leave the calculation to you as to how many photons of audio frequency per sample are running down the wires to the loudspeaker. But it is the other extreme we are interested in today, and I have some plots to show you precisely at what temperatures for a given frequency, or vice versa, one encounters quantum as opposed to classical behavior.

Let me remind you that by antenna temperature we mean that physical temperature of a matched resistor replacing the antenna that would produce the same intensity of radiation in the same frequency interval. Also that our interest is to see at what antenna temperatures non-classical behavior might manifest itself. The plots show both the black body spectrum at a given temperature and the density of photons per sample as a function of frequency at that temperature. Figure 33–4 shows you the number of photons which you can expect in a sample, as a function of frequency, when observing the cosmic microwave background.

I have chosen to start with this particular radiation to make the point that the temperature it produces in any and every antenna, will be the same. The independence of antenna temperature on the size of a telescope immersed in a black body is analogous to the energy per sample being independent of the width of the passband for broad band radiation. Both are related in a fundamental way to the Uncertainty Principle which says that a sample is a cell in (6-dimensional) phase space whose shape can be changed but whose volume is fixed and equal to Planck's constant cubed. Its occupancy therefore is determined only by the density of radiation in the appropriate part of phase space, which in turn is determined by the temperature characterizing it and the frequency of observation.

The microwave background is the only radiation that can fill the beam of any telescope whatever its size. In the case of sources that are smaller than the antenna beam, we will have so-called beam dilution resulting in an antenna temperature that can be much smaller than their brightness temperatures. Nevertheless, the antenna temperature is the only relevant one determining the density of radiation coming from the feed of the antenna to the receiver. Figure

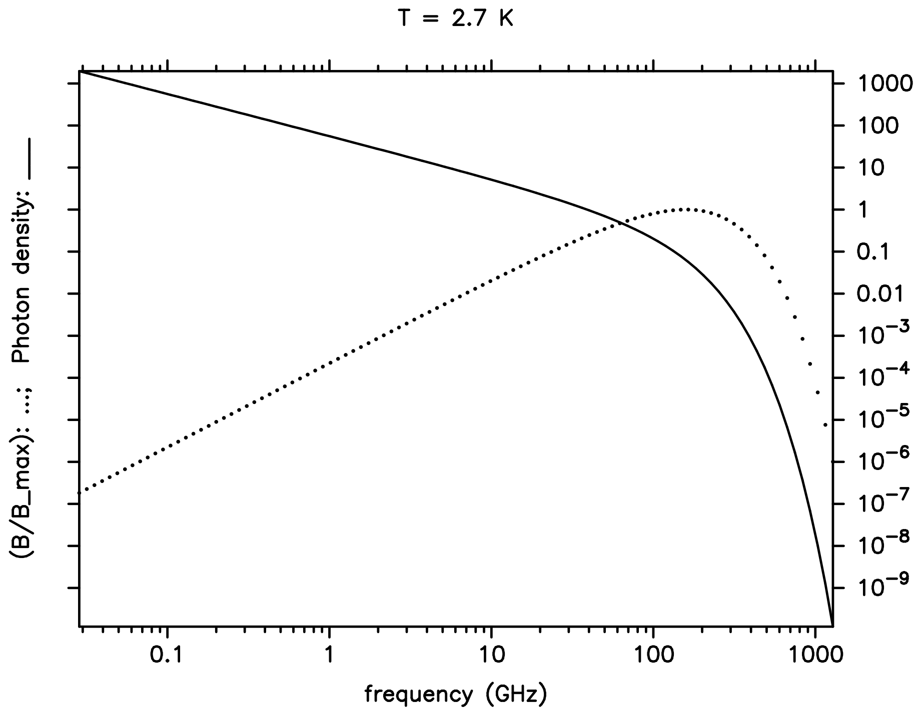


Figure 33–4. The microwave background radiation. The dotted line represents the intensity of a black body of 2.7 K, and the solid line, the number of photons per cell in phase space.

33–5 corresponds to a temperature of 100 K, typically the temperature of neutral interstellar gas. The only change from the previous figure is the relabeling of the frequency axis. The transition region between classical and quantum behavior has moved up as a consequence, and at the frequency of the hydrogen line of 1.4 GHz, a signal becomes sparse at antenna temperatures below a tenth of a degree Kelvin. At higher and lower frequencies the transition will occur at proportionately higher and lower antenna temperatures.

12. The Strength of Astronomical Signals

It is time to see whether all this discussion about weak and strong signals has any relevance to real life radio astronomy, say as practised here. The antennas in both arrays are of 25 meters diameter and would require about 10 Janskys of flux to produce one degree of antenna temperature. Relating this to what I have just said above, the dividing line between weak and strong signals in very very round numbers is at one Jansky per gigaHertz. At the lowest frequencies of operation there will be many situations where the signal can be considered highly classical. But for most of the sensitive observations made with the telescopes, particularly at high frequencies, the antennas collect less than one photon per sample on the average from the source! How come no attention whatever is paid to this circumstance, and no dire consequences result from making all observations in

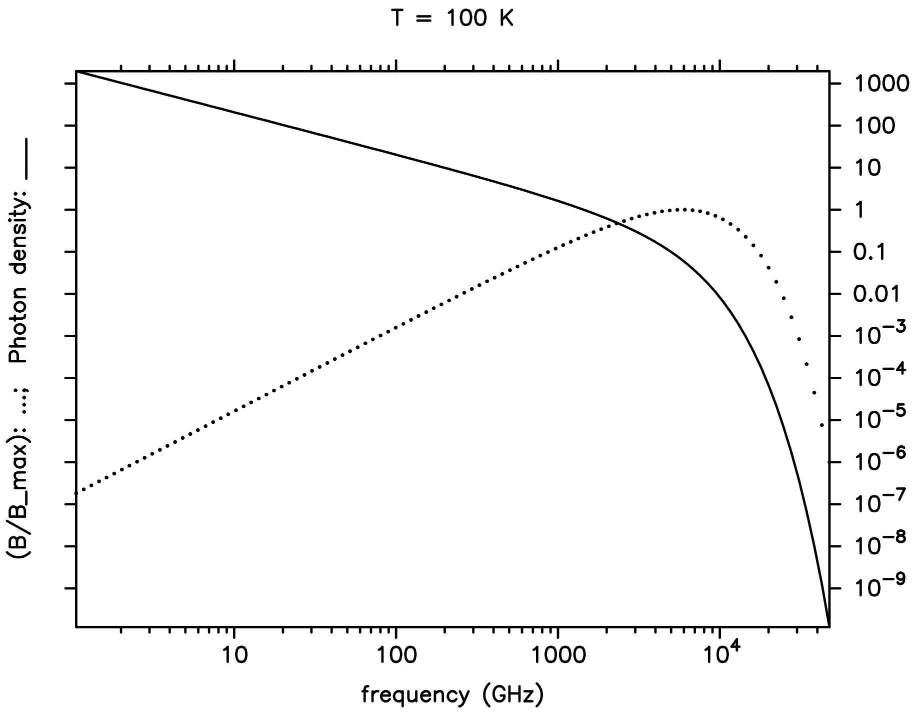


Figure 33–5. The radiation from a black body at 100 K. As in figure 33–4 the dotted line represents the intensity, and the solid line, the number of photons per cell.

the same standard fashion as if they were all of strong classical signals? To understand this we must first look at how astronomical signals are detected.

Barring radiation from our Sun, there are no signals from any astronomical sources that are strong enough to operate a measurement or recording device (other than the eye or a photographic plate) without our providing additional energy. Such devices are called detectors or amplifiers, most or all of which work by accelerating electrons that have either been liberated or set in motion by the astronomical signal. Photons of optical or higher frequencies have adequate energy to overcome the work function in many substances, and a measurement on the electron(s) so liberated is the best way to quantify such signals. In such detectors the photon is generally annihilated and the only information gathered is the energy of the photon, and its direction of arrival to the accuracy the telescope permits. The notion of phase of the signal in these cases is non-existent as already explained.

At usual radio frequencies the photons do not have enough energy to unbind and release electrons, but they can set them in motion in the conductors of which the antennas and feeds are made. But as these currents are orders of magnitude too weak to operate any devices, they have to be amplified first. The devices used for this purpose are called coherent amplifiers and are believed to produce strengthened versions of the input voltage or current waveforms. The amplifier in your hi-fidelity system is the archetypical example of such a device, the very name announcing (at a high decibel level) that the output signal is a faithfully amplified version of the input. We shall discuss in a moment the sustainability of this claim.

I would like to return briefly to the two-slit paradox which was dismissed rather rudely without stating clearly why it would be futile to investigate "which hole the photon went through". It is because any scheme to do so would necessarily introduce attenuation in one or both paths and modify the total diffraction pattern. The interference part of it will be degraded to the same degree as any information is obtained about passage through one of the two holes. Now if there existed truly a way to obtain two or more identical copies of an input signal, then we could have our cake and eat it too. If faithful amplification followed by precise division could provide identical copies, one of the pair could be used to measure what arrives at each hole, and the other to produce the interference pattern. I have already dwelt at length on the errors introduced when dividing a signal into two. To add to that difficulty, Heisenberg has ensured our failure by also prohibiting exact multiplication.

13. Coherent Amplifiers

The uncertainty associated with the energy-time pair of variables manifests itself as a minimum noise of one photon per sample at the input to any coherent amplifier with a large gain factor. Equivalent ways of thinking about this is in terms of the fluctuations of the vacuum field, or spontaneous emission which adds to the emission stimulated by the input signal. The net result is the prediction of a minimum theoretical noise temperature for any high-gain amplifier of order $h\nu/k_B$ degrees Kelvin (equation 28-15). For an input signal with a very large number of photons per sample, the addition of one more would affect neither amplitude nor phase. Coherent amplification therefore works best when the signal is already so strong that division would make no difference, like education being most effective when it is not needed!

On the other hand, compared to typical optical signals, one photon per sample is an enormous amount and is a thousand times as strong as a very bright star seen with a big optical telescope. It is in fact the reason why amplifiers are not used for astronomy at such frequencies and higher. To sum up, the points to remember about coherent amplification are the following. Very low density signals get badly corrupted even with ideal amplifiers. To obtain amplification without damage, the input has already to be a classical signal. No matter how weak the input, even if nothing, the output of a high gain amplifier is always of classical strength, permitting duplication or further amplification (of whatever came out) without further damage.

Historically, radio astronomy was created by radio amateurs and radar engineers who knew nothing about astronomy but all about amplifiers. They knew that their amplifiers added unwanted noise, and a lot of it, and the history of the field is closely tied to the saga of the development of better and lower noise receivers. The excitement of building newer types of receivers and experimenting with them was itself comparable to the astronomical rewards from the higher sensitivities attained. Clever circuits for vacuum tubes were followed by traveling wave amplifiers, parametric amplifiers, masers, field effect transistors and presently higher electron mobility versions of these. At the highest frequencies,

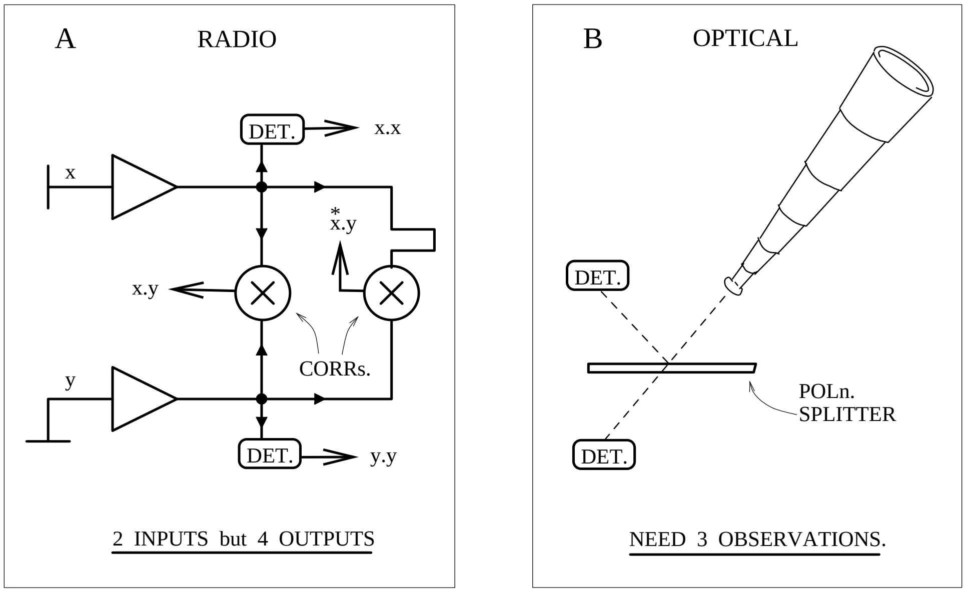


Figure 33–6. The measurement of polarization with and without amplifiers. A) At radio frequencies amplifying and splitting the signal from orthogonally polarized feeds permits the simultaneous determination of all four Stokes parameters. B) An optical telescope would require three observations in sequence for the same determination. Two of the six measurements obtained with different polarization splitters are redundant, but cannot be avoided.

using superconducting devices, one is approaching the achievement of minimum theoretical noise. But if a radio signal is as weak as the optical signals mentioned above, amplification must add an enormous amount of noise that will have to be integrated away before we can see the signal. Why then do we continue to amplify in radio astronomy? There are several reasons as I understand it.

The most serious, as already mentioned, is the work function as compared to the photon energies. Another is the temperature of our physical environment - the night is not dark in the radio. Quite apart from ground radiation which gets into every telescope and can be substantial, even the feeble microwave background of less than 3° K ensures that the occupation number at most radio frequencies is already high. In other words, even though the particular contribution to the signal that we seek is very very weak, it is already in a classical sea of noise and if there are benefits to be derived from retaining the associated aspects, we would be foolish to pass them up. One of them is the ability to measure phase. Even though the contribution of the desired signal to the measured phase is minute, it can be recovered with enough integration. And the most important of these benefits is the ability to duplicate ad nauseum once receiver noise has been added and a high multiplication factor obtained.

14. Epilogue

As a direct consequence of the above, radio astronomers routinely do many things their optical colleagues could not dream of doing. Let me mention some

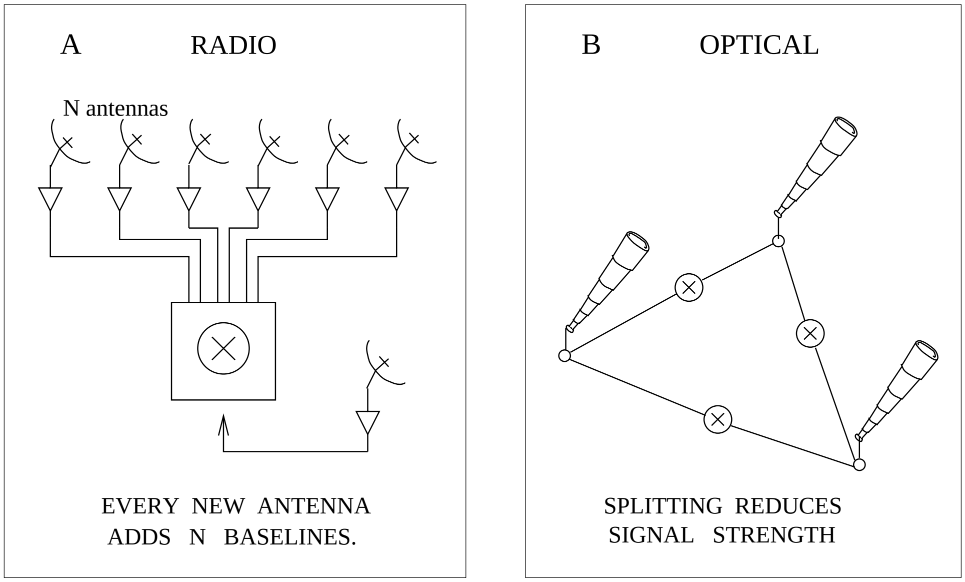

A
RADIO
N antennas
EVERY NEW ANTENNA ADDS N BASELINES.
B
OPTICAL
SPLITTING REDUCES SIGNAL STRENGTH

Figure 33–7. Aperture synthesis with and without amplifiers. A) In a radio array, the signal from each antenna is amplified and split N ways to be correlated simultaneously with the signals from all the other antennas. B) In an optical array of Michelson interferometers, splitting the signal reduces its strength and has to be compensated by increased observation to provide the same number of photons per baseline as without splitting.

examples, the first being the measurement of polarization of an astronomical signal. Amplifying and splitting the signal from orthogonally polarized feeds permits the simultaneous measurement of all four Stokes parameters, Figure 33–6A. An optical measurement would require three observations in sequence to obtain the same information because splitting the signal would worsen the signal/noise ratio, Figure 33–6B.

The second example is of a far more important advantage without which this school could not have been held. The signal from each of the twenty seven antennas of the array is amplified and split N ways to perform at the same time all the correlations with the signals from all the other antennas, Figure 33–7A. In the optical Michelson interferometers now being operated in several places, any splitting of the signal for other advantages like obtaining closure phase, Figure 33–7B, has to be compensated by further observations to accumulate the same number of photons as without splitting.

But the most spectacular achievement made possible by coherent amplification is of course VLBI where high fidelity recording and post facto reproduction is its very basis. And this technique has made another giant leap not so long ago with space VLBI in which the NRAO arrays play an important part. Optical technology has made fantastic progress spurred on by the communications industry and it is already many years since phase-locked optical oscillators became commonplace. But I have difficulty even imagining such a thing as optical VLBI. Maybe it will come, but for the moment it is for the radio astronomers to make hay while the sun shines on them as THE experts in synthesis imaging.

15. Acknowledgements

I am grateful to colleagues in several continents for encouragement, help, and suggestions for improvement of this printed version.

Synthesis Imaging in Radio Astronomy II
ASP Conference Series, Vol. 180, 1999
G. B. Taylor, C. L. Carilli, and R. A. Perley (eds.)

Index

Synthesis Imaging in
Radio Astronomy II

Designed by Greg Taylor
from series style
with help from Anton Zensus

Composed by the editors
with LaTeX by Leslie Lamport
based on TeX by Donald Knuth
in Computer Modern Times Roman

With LaTeX extensions
Fancyheadings
by Piet van Oostrum
Epsf by Tomas Rokicki
Psfig by Trevor Darrell

Index created with
MakeIndex by Pehong Chen
and Michael Harrison

Camera copy
with Dvips by Tomas Rokicki
at 600 dots per inch