TABLE OF CONTENTS

PREFACE

The Summer School on *Synthesis Imaging* held in Socorro, New Mexico from June 15 to June 22, 1988 was the third in which staff of the National Radio Astronomy Observatory prepared a series of lectures for serious students of synthesis imaging and image processing. This volume contains the edited texts of lectures from the series.

Purpose of the course.

The NRAO operates one 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. The observatory is also building the Very Long Baseline Array (VLBA), a synthesis telescope whose elements will be distributed from St. Croix to Hawaii and will be operated from Socorro. The major goal of this course, like that of its predecessors in 1982 and 1985, 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 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.

This 1988 version of the course saw our first attempt to divide the material into two main parts, which were given in successive weeks, with a "hands on" demonstration at the VLA during the intervening weekend.

The first segment of the course contains eleven lectures that describe the fundamentals of synthesis imaging and which could be read as a stand-alone course by the beginning student. The introductory lecture, by Barry Clark, sets the stage by considering the topic from the standpoint of physical optics. It treats the topic as that of measuring the coherence of a wave field, without regard to the practical details of correlation, or of the changing geometry of the problem as the Earth rotates. The second lecture, by Dick Thompson, starts from the viewpoint of the engineer who designs a practical synthesis array by considering the time-variable output of an interferometer located on the rotating Earth. Lecture 3, by Peter Napier, is new for this offering of the course. It describes those properties of the antennas that affect all that follows in the data processing. Lecture 4, by Larry D'Addario, discusses digital correlation, with new emphasis on the designer's choice between the 'XF' correlator as used in the VLA and the 'FX' correlator chosen for the VLBA. Lecture 5, on calibration and editing, has been substantially rewritten and was given this time by Ed Fomalont. The next six lectures are little altered from the 1985 school, except for minor additions and editorial changes. Lecture 6, given by Dick Sramek, discusses the techniques for forming the "dirty" image from the calibrated visibility data. This lecture provides the basis for all the subsequent discussions. Lecture 7, given by Pat Crane, discusses the sensitivity of the final image to various kinds of structure, with explicit application to the VLA. Lecture 8, given by Robert Braun, discusses how various deconvolution techniques can be used to minimize the imperfections that result from incomplete sampling. Lecture 9, given by Ed Fomalont, describes how the information in the image itself may sometimes be used to correct for distortions introduced by fluctuations in the instrument or in the atmosphere over it. Lecture 10, written by Ron Ekers but given in 1988 by Patrick Leahy, discusses how to recognize and correct the most common types of error that afflict synthesis images. Lecture 11, by Ed Fomalont, discusses how to extract astronomically interesting information from the final images.

The rest of the course is devoted to more advanced topics. Lecture 12, by Bill Cotton, sets the stage by discussing various ways in which the simplifying assumptions made in Lecture 1 are violated in common practical cases. A group of three new lectures then discusses topics in wide field imaging. Lecture 13, by Alan Bridle and Fred Schwab, discusses bandwidth and time-average smearing, expanding on an earlier treatment by Dick Thompson in the 1982 school. Lecture 14, by Rick Perley, treats the general three-dimensional problem of imaging with non-coplanar arrays. Lecture 15, by Tim Cornwell, discusses the techniques that can be used to make images that extend beyond the field of view of the primary interferometer elements. Lecture 16, by Rick Perley, is a substantially updated account of techniques for imaging faint emission in the presence of much stronger sources.

The expanded format of the 1988 course allowed us to double the time spent on two specialties that are of rapidly growing interest to the NRAO's user community—spectral "line" imaging and very-long-baseline interferometry. Lecture 17 contains a new introduction to spectral line imaging by Pjotr Roelfsema. Lecture 18, given by Jacqueline van Gorkom, contains a discussion of more advanced aspects of synthesis spectroscopy. Lecture 19, by Craig Walker, describes the basic principles and practice of VLBI, and Lecture 20, by Phil Diamond, describes the special problems of spectroscopic VLBI.

The next three lectures are new contributions that discuss imaging of sources with unusual properties. Lecture 21, by Tim Bastian, relates the special difficulties encountered when imaging the Sun with the VLA. Lectures 22 and 23, given by K. R. Anantharamaiah, discuss the unusual, but fascinating, problems associated with imaging spatially coherent emission and very bright sources.

The last two lectures deal with complementary aspects of the interaction between observers' needs and array design. Lecture 24, by Alan Bridle, addresses how to plan continuum and spectral line synthesis observing around given hardware capabilities; its Appendix is a primer for potential users of the VLA. Lecture 25, by Bob Hjellming, discusses how to design a synthesis array for a given class of observational problems; it emphasizes issues that affect the design of arrays for work at millimeter wavelengths.

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, all the references cited in this volume are listed in a section at the end.

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 Sandra Montoya, 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.

Acknowledgments.

We (especially the one of us, Rick Perley, who was in charge of local arrangements) thank Terry Romero and Mark Dinius for their invaluable assistance with the smooth running of the live school. We also thank Michael Andreatta and George Kessler for their work on the graphics.

The document has been typeset with $\mathcal{AMS}$-TeX, the American Mathematical Society's TeX macro system. We thank the AMS for permission to use this system and applaud Donald E. Knuth for inventing the TeX software.

Richard A. Perley
Frederic R. Schwab
Alan H. Bridle

OPENING REMARKS

Welcome to the Macey Center, on the Campus of the New Mexico Institute of Mining and Technology in Socorro, for the third NRAO summer course on Synthesis Imaging. The previous courses were held in the summers of 1982 and 1985, so it appears that a three-year cycle is developing.

In his introduction to the two previous volumes based on our earlier summer schools, Ron Ekers points out that many of the usual radio synthesis imaging techniques such as 'CLEAN', 'MEM' and 'Self-calibration' do not appear in standard textbooks. Fortunately, this situation is now remedied with the appearance of the second edition of *Radio Telescopes* by W. N. Christiansen and J. A. Högbom in 1985 and the new text *Interferometry and Synthesis in Radio Astronomy* by A. Richard Thompson, James M. Moran and George W. Swenson, Jr., published in 1986.

Why do we continue to feel that a summer course in synthesis techniques is required? The VLA has over 700 users per year. Many astronomers are relatively inexperienced in the area of synthesis imaging. Some are graduate students, and a few are even undergraduates. We also often have new users who are astronomers from other areas of specialization. Thus a synthesis summer school seems like a good idea. We are attempting to educate prospective and current users. Often we have heard the criticism that the VLA is a 'black box'. Certainly one of the major advances of the VLA has been that synthesis imaging has become possible for a large number of non-'black belt' radio astronomers. However, the use of this national facility has probably deprived many users from getting hands-on experience. One of the major aims of this summer school is to correct this situation.

There are many synthesis imaging radio telescopes in existence. As an example, the Australia Telescope in Culgoora, New South Wales will come into operation in the near future. The VLA has been in full operation since 1981. The Very Long Baseline Array of the NRAO will come into operation in the early 1990s. The intention in these lectures will be to provide a general discussion which will apply to all such synthesis instruments. Of course, most of the examples will be taken from the VLA experience.

All of you in attendance should recognize that this course would not have been possible without lots of hard work from many people. The most important is, of course, Rick Perley. Rick's enthusiasm and perseverance have made the 1988 workshop possible. Rick has had lots of support from Mark Dinius and Teresa Romero.

On behalf of the NRAO let me welcome you to Socorro. The staff of the NRAO is ready to help in the following week. Please call on us for any assistance. In the future we may perhaps look forward to the 1991 workshop being held in the new Array Operations Center of the VLA/VLBA.

W. Miller Goss

1. COHERENCE IN RADIO ASTRONOMY

Barry G. Clark

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 and 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 and 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 call $\mathbf{E}_\nu(\mathbf{R})$, are called the *quasi-monochromatic components* of the electric field

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

$\mathbf{E}(\mathbf{R}, t)$. 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.

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.

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 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 $V_\nu(\mathbf{r}_1, \mathbf{r}_2) = \langle E_\nu(\mathbf{r}_1) E_\nu^*(\mathbf{r}_2) \rangle$. 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:

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

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:

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

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:

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

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 analyses 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

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

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

Christiansen, W. N. and 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. and Mathur, N. C. (1968), "The interferometer in radio astronomy", *Proc. IEEE*, **56**, 2114–2130.

Yen, J. L. (1985), "Image reconstruction in synthesis radio telescope arrays", in *Array Signal Processing*, S. Haykin, Ed., Prentice–Hall (Englewood Cliffs, NJ), pp. 293–350.

2. THE INTERFEROMETER IN PRACTICE

A. Richard Thompson

In the first lecture it was shown that radio source imaging can be achieved by measuring the mutual coherence function of the electric fields at pairs of points in a plane normal to the direction of 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, using the u-v coordinate system introduced in Lecture 1 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 Lecture 1, from a discussion of the properties of the electric field, will emerge again as expressions for the response of an interferometer. Other discussions of the interferometer response can be found in Swenson and Mathur (1968), Fomalont (1973), Fomalont and Wright (1974), Meeks (1976), Christiansen and Högbom (1985), and Rohlfs (1986); a detailed and extensive review is given by Thompson, Moran and Swenson (1986).

1. 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

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

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.

I 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 Lecture 5. 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

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

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

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

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

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 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

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.

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

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

I 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 ($\mathrm{W\,m^{-2}\,Hz^{-1}}$). 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 brightness distribution $\mathcal{A}(\boldsymbol{\sigma})I(\boldsymbol{\sigma})$. By separating the real and imaginary parts of V in Equation 2–7 we obtain

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

and

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

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

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

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

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.

2. 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,

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

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

$$\begin{aligned} r &= A_0 |V| \int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2} \cos(2\pi\nu\tau_g - \phi_V)\, d\nu \\ &= A_0 |V| \Delta\nu \frac{\sin \pi\Delta\nu\tau_g}{\pi\Delta\nu\tau_g} \cos(2\pi\nu_0\tau_g - \phi_V) , \end{aligned} \tag{2–12}$$

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

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.

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

$$\frac{\sin \pi \Delta\nu \tau_g}{\pi \Delta\nu \tau_g} \approx 1 - \frac{(\pi \Delta\nu \tau_g)^2}{6} > 0.99\,, \tag{2–13}$$

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 Figure 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.

3. 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 $\nu_{\rm RF}$ with a local oscillator at frequency $\nu_{\rm LO}$ 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 $\nu_{\rm RF}$ and remains fixed when the observing frequency is changed. The signals at the frequencies $\nu_{\rm RF}$ and $\nu_{\rm LO}$ 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 $\nu_{\rm IF}$ is related to the mixer input frequencies by

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

Note that $\nu_{\rm LO}$ is a single-valued frequency, but $\nu_{\rm RF}$ and $\nu_{\rm IF}$ 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

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

where $\phi_{\rm LO}$ 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 $d\nu_{\rm IF}$, and integrating with respect to $\nu_{\rm IF}$ from $\nu_{{\rm IF}_0} - \Delta\nu/2$ to $\nu_{{\rm IF}_0} + \Delta\nu/2$. Thus:

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

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 $\nu_{\rm LO}$ 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:

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

whence

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

Here the differences in the signs of the various terms compared with those in Equations 2–15 occur because in lower sideband conversion a change in phase

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

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 Equations 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, and the greatest sensitivity is obtained by connecting the antenna directly to the mixer input without a filter to select only one sideband. The result is a double-sideband system, and the response is obtained from the sum of Equations 2–16 and 2–18:

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

Note that the delay-tracking error $\Delta\tau$ does not affect the phase of the cosine fringe term as it does in Equations 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 and Swenson (1986).

4. Fringe Rotation and Complex Correlators

The output from the correlator represented by Equation 2–16, 2–18 or 2–19 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 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,

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.

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 Equations 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 2–6 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 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

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.

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 7 for an analysis of signal-to-noise ratios.

5. 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 and Clark (1978).

6. 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. 2–6 and 2–7) become

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

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 I 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 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

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.

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

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

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

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

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) 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

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 ($\delta = 90°$).

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

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

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 $\theta_{\rm HPBW} = 1''$, $\theta_{\rm F} < 2\overset{\prime}{.}5$. 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 14.

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.

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

In two-element interferometers it is sometimes convenient to specify the baseline vector in terms of its length and the hour-angle and declination of the baseline direction on the Northern celestial hemisphere; see, for example, Rowson (1963). When a greater number of antennas are involved it is more 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, $(h = -6^{\mathrm{h}}, \delta = 0)$ 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

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

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 and 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:

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

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 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.

8. 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 I 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 usually not 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

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

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.

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.

9. 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.

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.

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 and 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 and 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

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.

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 proposed locations for antennas in the Very Long Baseline Array (VLBA) (Kellermann and Thompson, 1985), and examples of transfer functions, are shown in Figure 2–15. The effect of the addition of an antenna in low Earth orbit to an array like the VLBA is shown in Figure 2–16. The orbital motion fills out and extends the coverage very effectively. 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).

Figure 2–14. **(a)** The configuration of the 27 antennas of the VLA. **(b)** The transfer functions for four declinations with observing durations of $\pm 4^{\mathrm{h}}$ for $\delta = 0°$ and $45°$, $\pm 3^{\mathrm{h}}$ for $\delta = -30°$, and $\pm 5^{\mathrm{m}}$ for the snapshot. [From Napier, Thompson and 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. Examples of the *u*-*v* coverage obtained using a VLBI array similar to that of Figure 2–15a and one additional antenna in low Earth orbit. [From Preston *et al.* 1983.]

10. 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 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 and D'Addario 1982), as will now be described.

In observing continuum radiation one is interested in the mean brightness

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.

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

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

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

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

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.

11. 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

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 $\omega_e \tau_a$.

the average value for each interval is retained. In the subsequent processing the averaged visibility samples are assigned (u, v) values corresponding to the midpoints 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 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 $\operatorname{cosec}\delta$ 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), "Nouveau réseau pour l'observation radioastronomique de la brillance sur la Soleil à 9530 Mc/s", *C. R. Acad. Sci.*, **240**, 942–945.

Bracewell, R. N. (1966), "Optimum spacings for radio telescopes with unfilled apertures", 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., and Thompson, A. R. (1973), "The Stanford five-element radio telescope", *Proc. IEEE*, **9**, 1249–1257.

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

Fomalont, E. B. (1973), "Earth-rotation aperture synthesis", *Proc. IEEE*, **61**, 1211–1218.

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

Frater, R. H. (1984), "The Australia Telescope—the radio astronomer's highway to the future", *Proc. Astron. Soc. Australia*, **5**, 440–445.

Granlund, J., Thompson, A. R., and Clark, B. G. (1978), "An application of Walsh functions in radio astronomy instrumentation", *IEEE Trans. Electromag. Compat.*, **EMC-20**, 451–453.

Högbom, J. A. and Brouw, W. N. (1974), "The synthesis radio telescope at Westerbork. Principles of operation, performance and data reduction", *Astron. Astrophys.*, **33**, 289–301.

Kellermann, K. I. and Thompson, A. R. (1985), "The Very Long Baseline Array", *Science*, **229**, 123–130.

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

Michelson, A. A. (1890), "On the application of interference methods to astronomical measurements", *Phil. Mag.*, Ser. 5, **30**, 1–21.

Moffet, A. T. (1968), "Minimum redundancy linear arrays", *IEEE Trans. Antennas Propagat.*, **AP–16**, 172–175.

Napier, P. J., Thompson, A. R., and Ekers, R. D. (1983), "The Very Large Array: design and performance of a modern synthesis radio telescope", *Proc. IEEE*, **71**, 1295–1320.

Preston, R. A., Burke, B. F., Doxsey, R., Jordan, J. F., Morgan, S. H., Roberts, D. H., and Shapiro, I. I. (1983), "The future of VLBI operations in space", 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.

Rowson, B. (1963), "High resolution observations with a tracking radio interferometer", *Mon. Not. Roy. Astron. Soc.*, **125**, 177–188.

Ryle, M. (1952), "A new radio interferometer and its application to the observation of weak radio stars", *Proc. Roy. Soc.*, **211A**, 351–375.

Ryle, M. (1962), "The new Cambridge radio telescope", *Nature*, **194**, 517–518.

Ryle, M. (1972), "The 5-km radio telescope at Cambridge", *Nature*, **239**, 435–438.

Swenson, G. W., Jr. and Mathur, N. C. (1968), "The interferometer in radio astronomy", *Proc. IEEE*, **56**, 2114–2130.

Thompson, A. R., Clark, B. G., Wade, C. M., and Napier, P. J. (1980), "The Very Large Array", *Astrophys. J. Suppl. Ser.*, **44**, 151–167.

Thompson, A. R. and D'Addario, L. R. (1982), "Frequency response of a synthesis array: performance limitations and design tolerances", *Radio Science*, **17**, 357–369.

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

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

3. THE PRIMARY ANTENNA ELEMENTS

Peter J. Napier

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, Love 1978; Rudge *et al.* 1982) 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 ν_{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 included in Lecture 7. 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 and Ekers (1983) and Thompson, Moran and 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 synthesized image. Derivation of these expressions can be found in standard textbooks (e.g., Kraus 1986, Chapter 3).

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.

Here, as in Lecture 2, we require a notion 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 Fig. 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 Relationship 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, Relationship 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, $f(x, y)$, in the aperture of an antenna and the complex far-field voltage radiation pattern, $F(u, v)$, of the antenna (Kraus 1986, Section 6–8):

$$F(u, v) = \iint_{\text{aperture}} f(x, y) e^{2\pi i(ux+vy)}\, dx\, dy \,, \tag{3–6}$$

and

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(u, v) e^{-2\pi i(ux+vy)}\, du\, dv \,. \tag{3–7}$$

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

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

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,[1] 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

[1] 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, $f(x)$, and the radiation pattern, $F(u)$, are shown for a one-dimensional example. In general both quantities are complex. Only the amplitudes are shown here.

sidelobes, but the radiation pattern does not have to be Hermitian (i.e., need not be conjugate symmetric).

The form of $f(x, y)$ for an antenna is determined by the way in which the antenna feed illuminates the aperture. In general, the more that $f(x, y)$ 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 $f(x, y)$, and their $F(u, v)$, 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 ($f(x, y)$ = constant) over the whole aperture. In this case, for a circularly symmetric aperture of diameter D, $F(u) = J_1(\pi D u)/u$, 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 discussed in Section 4.1.2.

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, 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

Figure 3–4. Equatorial and Altazimuth antenna mounts.

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 and 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.

3.2.1. Choice of mount.

The choice of the mount is usually between an equatorial mount and an elevation-over-azimuth 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

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.

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, Moran and 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 order 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 polarized 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 only about this axis is required 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.

3.2.2. 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 3.1.3.

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 (see Section 3 of Lecture 7). The aper-

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.

ture efficiency, η, is the product of a number of different loss factors,

$$\eta = \eta_{\mathrm{sf}}\, \eta_{\mathrm{bl}}\, \eta_s\, \eta_t\, \eta_{\mathrm{misc}}\ , \tag{3–9}$$

where η_{sf} = reflector surface efficiency, η_{bl} = reflector blockage efficiency, η_s = feed spillover efficiency, η_t = illumination taper efficiency, and η_{misc} = miscellaneous efficiency losses due to reflector diffraction, feed position phase errors, and feed match and loss.

4.1.1. 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

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

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 Expression 3–10 (Ruze 1966). Expression 3–10 predicts that for an r.m.s. error of $\lambda/16$, $\eta_{\mathrm{sf}} = 0.54$, 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 η_{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 η_{sf} performance requires careful structural design for wind, thermal and gravitational loading, together with precise reflector panels and an accurate panel setting technique.

4.1.2. 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 $0.75 < \eta_{\mathrm{bl}} < 0.95$. η_{bl} is given (Rudge *et al.* 1982, p. 179) by

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

The effective blocked area is the blocked area weighted for the illumination taper in the aperture. Equation 3–11 shows, for small blockage, that the loss

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.]

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

4.1.3. 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 Fig. 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

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.

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 a 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 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.

4.1.4. 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

$$\eta_t = \frac{\left(\iint_{\text{aperture}} |f(x, y)| \, dx \, dy \right)^2}{A \iint_{\text{aperture}} |f(x, y)|^2 \, dx \, dy} . \tag{3–13}$$

η_t equals 1 for uniform illumination ($f(x, y) \equiv 1$). 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 Fig. 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

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

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 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.

4.1.5. 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 wavelengths.

Table 3–1.

λ	$\eta_{\rm sf}$	$\eta_{\rm bl}$	η_s	η_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 and Ryle 1977; Mayer *et al.* 1983) and is ideally suited for use on the primary elements of a synthesis array. This measurement is based on the Fourier transform relationship of Equation 3–7. $F(u, v)$ 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. $f(x, y)$, the complex aperture distribution of the scanning antenna, is then computed using Equation 3–7. $|f(x, y)|$ then directly shows the aperture blockage and illumination, whilst $\arg\bigl(f(x, y)\bigr)$ provides information about reflector surface errors and feed location errors. Figure 3–11 shows the results of a holographic measurement of a VLA antenna. $|F(u, v)|$ shows the expected sidelobe enhancement in the planes of the subreflector support spars, and $|f(x, y)|$ 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. 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 $2\mathcal{A}\bigl(\theta_{3\mathrm{dB}}(\frac{1}{2} + \frac{1}{20})\bigr) \approx 0.87$, which will significantly reduce the accuracy of the outer part of the image.

The VLA antenna pointing accuracy is 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

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

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

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.

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 called the "instrumental polarization" or, in Lecture 5 (Sec. 9.2), the polarization "crosstalk", D. 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 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

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.

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, $f_c(x,y)$ has the antisymmetric pattern shown schematically in Figure 3–14(a). The average value of $f_c(x,y)$ is zero, so it does not give rise to on-axis instrumental polarization. The cross-polarized radiation pattern, $F_c(u,v)$, which is the Fourier transform of $f_c(x,y)$, 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 $F_c(u,v)$ for a VLA antenna is also shown in Figure 3–14. The effects of $F_c(u,v)$ 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 complicated on an alt–az antenna because $F_c(u,v)$ rotates with respect to the source, as explained in Section 3.2.1. 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 $F_c(u,v)$ 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

Figure 3–14. Cross-polarized beam of a reflector antenna. **(a)** Schematic diagram of $f_c(x, y)$ corresponding to the cross-polarized field in Figure 3–12. **(b)** Form of $F_c(u, v)$. **(c)** $F_c(u, v)$ 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.

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 and Turrin 1973).

References

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, 1982*, A. R. Thompson and L. R. D'Addario, Eds., NRAO (Green Bank, WV).

Chu, T. S. and Turrin, R. H. (1973), "Depolarization properties of offset reflector antennas", *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., and Welch, W. J. (1987), "Aperture efficiency enhancement in a Cassegrain system by means of a dielectric lens", *Radio Science*, **22**, 1091–1101.

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

Love, A. W., Ed. (1978), *Reflector Antennas*, IEEE Press, New York.

Mayer, C. E., Davis, J. H., Peters, W. L., and Vogel, W. J. (1983), "A holographic surface measurement of the Texas 3.9-m antenna at 86 GHz", *IEEE Trans. Instrum. Meas.*, **IM-32**, 102–109.

Napier, P. J., Thompson, A. R., and Ekers, R. D. (1983), "The Very Large Array: Design and performance of a modern synthesis radio telescope", *Proc. IEEE*, **71**, 1295–1320.

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

Ruze, J. (1966), "Antenna tolerance theory—A review", *Proc. IEEE*, **54**, 633–640.

Scott, P. F. and Ryle, M. (1972), "A rapid method for measuring the figure of a radio telescope reflector", *Mon. Not. Roy. Astron. Soc.*, **178**, 539–545.

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

4. CROSS CORRELATORS

Larry R. D'Addario

1. Introduction

This lecture will describe the operation of the central correlator of a synthesis telescope. I shall be more concerned here with details of the hardware than the earlier lecturers have been. In modern telescopes, major portions of the correlators—including delay lines, multipliers, and integrators—are implemented digitally; this is done for very good reasons, but it leads to results that are significantly different from what one would predict by analyzing a continuous-time, analog model. For this reason, I will concentrate on the digital implementation of correlators and pay considerable attention to the process of digitizing the signals from the antennas.

In addition, I will describe how a synthesis telescope can be used for spectroscopy; that is, how a correlator can provide visibility measurements as a function of frequency over the receiver passband. Spectral synthesis differs from continuum synthesis, and it also differs significantly from single antenna spectroscopy. Some of the differences will be pointed out here; Lectures 17 and 18 will consider these special problems in greater depth.

2. Correlators in General

The preceding lectures dealt mainly with the correlation of quasi-monochromatic signals. I would now like to generalize to the case of wide bandwidth signals; this leads naturally to an understanding of spectroscopic cross correlation. Sometimes the signal bandwidth is not quasi-monochromatic, but the main reason for considering wide bandwidth correlators is that the signals normally are converted to a low center frequency by the time they reach the correlator inputs. Their fractional bandwidth $\Delta\nu/\nu_0$ can then be very large.

The cross correlation function of two real signals $v_i(t)$ and $v_j(t)$ is

$$x_{ij}(\tau) = \langle v_i(t) v_j(t+\tau) \rangle \,, \tag{4–1}$$

where $\langle\cdot\rangle$ represents the ensemble (statistical) average. This is a real function of delay τ, and can be estimated by the simple correlator of Figure 4–1. In the special case that v_i and v_j are narrow-band signals centered at ν_0 with bandwidth $\Delta\nu \ll \nu_0$, it is clear that $x_{ij}(\tau)$ is nearly sinusoidal in τ, with period ν_0^{-1} (see Fig. 4–2). That is, we can write

$$x_{ij}(\tau) = x_R \cos 2\pi\nu_0(\tau - \tau_0) + x_I \sin 2\pi\nu_0(\tau - \tau_0) \,, \tag{4–2}$$

Figure 4–1. A simple (real) correlator.

for τ in the vicinity of reference delay τ_0.[1] Then $x_{ij}(\tau)$ is specified for a wide range of τ by the single complex number $R_{ij} = x_R + ix_I$. This defines the complex cross power for narrow-band signals, where the signals themselves are real functions of time. (Complex cross power becomes complex visibility after astronomical calibration.) It is thus not necessary to measure $x(\tau)$ for all τ, but only for two nearby values of τ. Convenient choices are τ_0 and $\tau_0 + \Delta\tau$, where $\Delta\tau = 1/(4\nu_0)$. This leads to the "complex correlator" of Figure 4–3.

If the signals are not narrow-band, then $x_{ij}(\tau)$ will not be sinusoidal, but the concept of complex cross power is still useful. We can imagine using a bank of filters to break up each wide band into many disjoint narrow bands, and then connecting each pair of outputs to a complex correlator, as in Figure 4–4. Here each box "CC" represents that part of Figure 4–3 within dashed lines, but each has a delay $\Delta\tau_k = 1/(4\nu_k)$ appropriate to its own frequency. If the variation of correlation with frequency is not of interest, such as in the case of a continuum source, we can add together all of the outputs; this leads to a more accurate measure of the average correlation over the full bandwidth. This sum, in the limit where the number of filters gets very large (so that the sum approaches the integral over frequency), we may now define to be the complex cross power $R_{ij}(\tau_0)$ in the general case of arbitrary bandwidth:

$$R_{ij}(\tau_0) = \lim_{K\to\infty} \sum_{k=1}^{K} x_{Rk} + ix_{Ik}\,. \tag{4–3}$$

Now, it turns out that this same quantity can be measured without the elaborate filtering and multiple correlators of Figure 4–4. It is merely necessary

[1] We assume that the correlation function has a well-defined maximum absolute value, and that this occurs very near τ_0. This will hold for most wide-bandwidth signals. The value of τ_0 depends of the physical situation; in an interferometer, it includes the path difference from the source to each antenna as well as the difference in transmission line lengths from each antenna to the correlator. It may be negative, and can be much larger than the reciprocal bandwidth.

Figure 4–2. Cross correlation function of quasi-monochromatic signals with rectangular passbands centered at ν_0. In this plot, $\Delta\nu/\nu_0 = 0.2$ for clarity, but often it is much smaller.

to replace the quarter-cycle delay in Figure 4–3 (box "H") with a filter that passes all frequencies but shifts the phase of each by $\pi/2$. For narrow bandwidths, this is the same thing as a quarter-cycle delay; for wide bandwidths, a more complicated filter is needed, but such filters can be built. To save time, I will not give the proof that this is the same as summing the outputs of the filter bank correlator, but perhaps you can see that it is plausible. Mathematically, the $\pi/2$ phase shift operation is equivalent to the Hilbert transform (also called the Kramers–Krönig transform by some physicists; see, e.g., Bracewell 1978 for properties of the Hilbert transform).

The preferred method of making a complex correlator for wide-band continuum observations is therefore that of Figure 4–3, where "H" is a Hilbert transform filter. But the filter bank correlator of Figure 4–4 would obviously be useful for spectroscopy, where we would record the output of each complex correlator separately, rather than adding them together.

A nearly equivalent way to obtain the spectroscopic measurements is illustrated in Figure 4–5. This machine measures the real cross correlation function at a large number of closely spaced delays near τ_0, and computes the discrete Fourier transform (DFT) of the result. It takes $2K$ samples of the correlation function to obtain the complex visibility at K frequencies. I will describe both types of spectroscopic correlator in more detail later.

The discussion so far has been rather heuristic, so I will now try to fill in some of the associated mathematics. The (real) correlation function of two arbitrary signals is defined by Equation 4–1. Now consider its (inverse) Fourier

Figure 4–3. A complex correlator. For narrow bandwidth signals, "H" is a quarter-cycle delay, $\Delta\tau = 1/4\nu_0$.

Figure 4–4. A wide-band complex correlator synthesized from narrow-band complex correlators, or a spectroscopic correlator. Each box labeled "CC" is as indicated in Figure 4–3.

transform

$$r_{ij}(\nu) = \int_{-\infty}^{\infty} x_{ij}(\tau)\, e^{-2\pi i\nu(\tau-\tau_0)}\, d\tau\,, \tag{4–4}$$

which is called the cross power spectrum. (Recall that, similarly, the inverse transform of the *auto*correlation function of a signal is the signal's power spectrum; but the latter is always real and non-negative, whereas cross power is generally complex.)[2] Next, the complex cross correlation function is defined as

$$R_{ij}(\tau) = 2\int_{0}^{\infty} r_{ij}(\nu)\, e^{+2\pi i\nu(\tau-\tau_0)}\, d\nu\,; \tag{4–5}$$

[2] In Equation 4–4, I have inserted a time-shift of τ_0 before transforming. This definition is convenient if x_{ij} peaks near τ_0, because then r_{ij} will be nearly constant.

Figure 4–5. A spectroscopic correlator with frequency analysis after correlation.

i.e., it is twice the Fourier transform of Equation 4–4 with negative frequencies deleted. Notice that the correlator of Figure 4–5 approximates the right-hand side of Equation 4–4, and that adding up the outputs approximates the r.h.s. of Equation 4–5 for $\tau = \tau_0$. More precisely, the operation of the spectroscopic correlator of Figure 4–5 is described by

$$\hat{x}_k = \sum_{l=0}^{2K-1} \left[\frac{1}{T} \int_0^T v_i(t - \tau_0 - l\delta\tau) v_j(t)\, dt \right] e^{-2\pi i l k/2K} . \tag{4–6}$$

The expression in brackets is the output of each simple correlator. Comparing Equations 4–4 and 4–6, we see that

$$r_{ij}(k/\delta\tau) \approx \hat{x}_k \delta\tau . \tag{4–7}$$

There is another way to compute the continuous cross correlation function:

$$\begin{aligned} R_{ij}(\tau) &= \tfrac{1}{2}\langle [v_i(t) + i\tilde{v}_i(t)]^* \, [v_j(t+\tau) + i\tilde{v}_j(t+\tau)] \rangle \\ &= \langle v_i(t)\, v_j(t+\tau) \rangle + i \langle \tilde{v}_i(t)^* \, v_j(t+\tau) \rangle , \end{aligned} \tag{4–8}$$

where $\tilde{v}$ represents the Hilbert transform of v. This follows from Equations 4–1, 4–4, and 4–5 (proof left as an exercise). Notice that Equation 4–8 describes the operation of the complex correlator of Figure 4–3.

3. Digital Implementations

Major portions of modern correlators are implemented digitally, for the following reasons: (1) digital operations are precisely defined and repeatable

Figure 4–6. An example of a quantizer transfer function (solid lines); this quantizer has seven levels. The dashed line is the line defined by $v_q = v$, and the difference between it and the transfer function is the quantization noise, δ.

(analog circuitry is subject to environmental conditions such as temperature and humidity); (2) digital circuits can be exactly replicated at low cost when many identical elements are needed; (3) for the long baselines ($> 10^4$ m for connected elements and $> 10^6$ m for VLBI) and wide bandwidths (10^8 Hz) now used, the delay lines must have a large ratio of length to resolution ($\gg L\Delta\nu/c > 10^4$), and only digital delay lines can do this with the necessary accuracy and stability.

3.1. Digitization.

The digital correlator must first convert the signals to digital form. This requires two distinct operations: sampling, which converts a continuous-time signal $v(t)$ to a discrete-time sequence of its samples $\{v(t_k) : k = 0, 1, \ldots\}$; and quantizing, which converts a continuously variable value to one of a finite set of values. This combination of a sampling device and a quantizing device is called a *digitizer*. For any finite length of time, the digitized signal can be represented by a finite number of bits and can be stored and processed with logic circuits. The signal can be sampled and then quantized, or quantized and then sampled, and the result will in principle be the same (as long as the circuits behave ideally).

If the signal $v(t)$ is strictly limited to frequencies between zero and $\Delta\nu$, then, according to the sampling theorem (Shannon, 1949), it is fully described by its samples taken at intervals $\Delta t \leq 1/(2\Delta\nu)$; that is, $v(t)$ can be exactly reconstructed from these samples. Strictly speaking, the sampling must go on for all time—but in practice it is only necessary to have a very large number of samples, and this condition is easily fulfilled in our case. Thus, sampling at the rate $2\Delta\nu$ (called the Nyquist rate), or faster, loses no information at all.

Quantization, however, does lose information. Consider Figure 4–6, which shows the transfer function of a typical quantizer. Here v is the quantizer's instantaneous input, and v_q is the corresponding output; this example shows seven distinct output states. Without loss of generality, the scale has been chosen so that v_q is the integer nearest the input value; then I can write $v_q = v + \delta$, so that the quantizer may be described as adding a signal δ to the input, sufficient to round it to the nearest integer. If the signal is random noise, then δ will also be noise-like, and for a reasonably chosen transfer function, δ will have zero mean. Thus, the quantizer can be viewed as adding noise to the signal. This is known as "quantization noise," and in a radio telescope it is the source of the degradation in signal-to-noise ratio associated with the use of digital correlators (the correlator efficiency η_c is used in Lecture 7).

Now imagine that the quantization is done before sampling. If the original signal has bandwidth $\Delta\nu$, then the quantized signal has a larger bandwidth (including harmonics), because the quantization noise $\delta(t)$ is not bandlimited. If we now sample at the rate $2\Delta\nu$, additional information is lost because the larger bandwidth is undersampled. This information can be partially recovered by sampling at a higher rate. Thus, it is not straightforward to apply the sampling theorem to signals that are also quantized, and the digitizer must be analyzed as a unit.

Nevertheless, if the signal consists of Gaussian noise, then even with Nyquist sampling very coarse quantization can be used with remarkably little loss of information. In synthesis telescopes, we are interested in the cross correlation function of two signals that are jointly Gaussian random processes. It can be shown (Van Vleck and Middleton, 1966; Cooper, 1970; Hagen and Farley, 1973) that the cross correlation function of digitized signals (for most reasonable quantizations) is a monotonic function of that of the original signals. However, a *measurement* of the digitized cross correlation in finite averaging time will have a larger relative variance than a similar measurement of the original signals, due to the quantization noise. Table 4–1 shows the resulting "loss" of signal-to-noise ratio for various cases, computed assuming rectangular power spectra of width $\Delta\nu$ and with the quantization levels optimized for each case (Hagen and Farley, 1973). Even the extreme case of two level quantization, where only the sign of the signal is retained, gives 64% of the undigitized signal-to-noise ratio. Two level quantization has been extensively used in VLBI, where the digitized signal is stored on tape, because it can be shown that this leads to nearly the maximum information per length of tape. Finer quantization or faster sampling gives higher sensitivity, at the cost of more complexity and more expensive components in the correlator. For the VLA, three level quantization was chosen as a reasonable compromise, with Nyquist sampling at the widest bandwidth (50 MHz) and up to four times Nyquist at some narrow bandwidths.

Besides sampling and quantizing, practical digitizers must do one more job: each quantized sample must be *encoded* digitally, typically as a binary number. For two level quantization, the obvious choice of one bit per sample is the only reasonable one. But with more levels, various encodings are possible, especially considering that the various levels do not occur with equal probability. If the signal is to be stored (say, on magnetic tape) or transmitted over an expensive channel before correlation, then it is important to choose a code that minimizes the number of bits needed. It turns out that, with optimum coding, the total

Table 4–1.
Signal-to-Noise Ratio vs. Quantization and Sampling Rate

Quantization	Sampling Rate	$\frac{\text{S/N (digital)}}{\text{S/N (continuous)}}$
2-level (1 bit)	$2\Delta\nu$	.64
	$4\Delta\nu$	.74
3-level	$2\Delta\nu$	.81*
	$4\Delta\nu$	.89
4-level	$2\Delta\nu$	.88
	$4\Delta\nu$	.94
∞-level (continuous)	$2\Delta\nu$	1.00
	$4\Delta\nu$	1.00

*VLA Case.
All cases assume rectangular bandpasses of width $\Delta\nu$, signal levels adjusted to maximize the signal-to-noise ratio, and small correlation coefficients.

number of bits needed to achieve a given sensitivity is minimized for three level quantization (D'Addario 1984); in this sense, three level quantization is optimum.

3.2. Quantization corrections.

As I mentioned, the cross correlation of digitized signals is a monotonic function of that of the original signals. By knowing this function, or rather its inverse, the desired cross correlation can be recovered. For example, with two level quantization it has been shown (Van Vleck and Middleton, 1966) that

$$x_{ij}(\tau) = \sigma_i \sigma_j \sin \frac{\pi \rho_{ij}(\tau)}{2}, \tag{4–9}$$

where ρ_{ij} is the correlation coefficient (normalized) of the digitized signals, and where $\sigma_i^2 = \langle v_i^2 \rangle$ and $\sigma_j^2 = \langle v_j^2 \rangle$ are the average power levels of the signals. Equation 4–9 is often called the "Van Vleck correction," after the author who first used it, although he did so in a much different context. Notice that, for two level quantization, the signal powers must be separately determined in order to get the cross power, since this information is completely lost in the quantization.

For three (or more) level quantization, the situation is more complicated. The correction function does not have a closed form expression, and it depends non-linearly on both the measured correlation coefficient ρ_{ij} and the signal powers. As an example, for the three level case I can write

$$\rho_{ij}(\tau) = f_3(x_{ij}(\tau); \sigma_i, \sigma_j), \tag{4–10}$$

where f_3 is an integral of the joint probability density function of the two signals. Then

$$x_{ij}(\tau) = f_3^{-1}(\rho_{ij}(\tau); \sigma_i, \sigma_j). \tag{4–11}$$

Once again, the signal powers are needed, but now they can be determined from the digitized signals themselves using digital autocorrelators. The form of f_3^{-1} can be assumed known, and can be calculated to any desired accuracy if an adequate computer is available (for numerical methods pertaining to this case, see Schwab, 1979, and D'Addario *et al.*, 1984). Similar concepts apply to other quantizations. Some examples are shown in Figure 4–7.

It is worth noting that relationships like Equations 4–9 and 4–11 do not depend on the sampling rate, the bandwidth, or the shape of the spectrum. However, all of these results apply only if the signals are zero mean, Gaussian random processes.

It turns out that if $\rho_{ij} \ll 1$, then x_{ij} is very nearly proportional to ρ_{ij} for all reasonable quantizations (see Fig. 4–7). This is apparent from Equation 4–9 in the two level case. We get $\rho_{ij} \ll 1$ when the antenna temperatures due to the source are much less than the system temperatures. Then a detailed computation of the correction can be avoided, provided that the signal powers σ_i^2 and σ_j^2 remain constant, because the proportionality factor drops out in astronomical calibration.

3.3. Gain corrections and ALC loops.

Notice that the cross power $x_{ij}(\tau)$ and its spectrum $r_{ij}(\nu)$ are referred to the correlator inputs. Actually, these quantities are not directly of interest; we would rather know the cross power spectrum of the signals received at the antennas. Denoting the latter by $r'_{ij}(\nu)$, we have

$$r_{ij}(\nu) = g_i(\nu) g_j^*(\nu) r'_{ij}(\nu + \nu_{\rm LO}), \tag{4–12}$$

Figure 4–7. Quantization correction functions for various quantizations. In each case the signal powers are set for maximum signal-to-noise ratio. The curves are labeled according to the number of quantization levels; 4a uses a simplified multiplier (see Cooper, 1970).

where $g_i(\nu)$ and $g_j(\nu)$ are the complex voltage gains of the signal paths from the antennas to the correlator, and $\nu_{\rm LO}$ is the net local oscillator frequency, accounting for all frequency conversions. If the gains are slowly varying, their effects can be largely accounted for by astronomical calibration (I will not discuss the details here, since Lecture 5 does so). However, in order to make life easier for electronics engineers, it often happens that no attempt is made to keep the gains constant (which would be hard); on the contrary, the gains are deliberately varied in order to keep the signal powers constant at the correlator (which is easier). This is done with automatic level control (ALC) loops.

For the programmer and the astronomer, ALC loops are a mixed blessing. They usually will cause the gains to change between the observation of a calibrator and that of a source being measured, either because one source is strong enough to contribute substantially to the total noise, or because the sources are in different parts of the sky, so that the noise contributions from the atmosphere and from ground radiation are different. The correlator has no way of knowing about this, since its input levels are constant; so an independent means of monitoring the gains must be provided. This is often done by adding a fixed, known signal to each receiver input and detecting it near the correlator input. A switched noise signal at each antenna and a synchronous, square law detector at each correlator input are effective in measuring the magnitudes of the gains; it is usually assumed that the phase is sufficiently stable. Given such measurements, the measured correlation function $x_{ij}(\tau)$ (after quantization correction) can be scaled in the computer to refer to the receiver inputs rather than to the correlator inputs. This is sometimes called the "system temperature correction," because, with ALC, the gain is inversely proportional to the system temperature. Notice that this scheme measures only the average gain across the passband, so the scaling will be strictly correct only if $g_i(\nu)$ is flat; and that the ratio of source to calibrator gain will be correct only if the gain changes by the same factor at all frequencies.

The following advantages of ALC loops often outweigh these difficulties: (a) changes in the gains of electronic components with time and temperature are cancelled (if they are the same at all frequencies); (b) the correlator input powers can be kept at the value that gives the best signal-to-noise ratio; and (c) the quantization correction calculations are simpler for constant input powers.

3.4. Digital circuits.

Figure 4–8 shows some details of a digital implementation of a simple correlator, including delay line, multiplier, and integrator. I include this mainly to give some feeling for the quantity of circuitry involved and the speeds at which it must operate. Generally, faster logic and memories take up more space, consume more power, and are more expensive than slower ones. High effective speeds can be achieved by having many slow circuits operating in parallel, and the various trade-offs often favor taking this approach. Thus, the delay line can be implemented mostly with slow memory, using small amounts of fast memory (shift registers) to buffer the input and output. Similarly, the integrator memory can be broken up into two or more stages, with slower devices used to accumulate for longer time periods. Multipliers, on the other hand, are generally operated at the full sampling rate; but since only two- or three-state signals usually need to be multiplied, the logic of a multiplier is quite simple.

When signals from a large array of antennas must be correlated, it usually turns out that the multipliers and integrators dominate the circuitry, since $N(N-1) \approx N^2$ of them are needed for complex cross correlation of N signals, whereas only $2N$ digitizers and delays are needed. They dominate even more in a spectroscopic correlator, where $KN(N-1)$ are needed for K frequencies, compared with only N digitizers and delays. Therefore, a design strategy that reduces the required number of multipliers is helpful. It turns out that it is sometimes possible to build multipliers and first-stage integrators that can operate much faster than the sampling rate; this is especially true when the receiver bandwidth is deliberately made small, so that a low sampling rate can be used. Then if a substantial number of samples can be stored temporarily in a buffer memory, the same multiplier/integrator can be time-shared among many correlators. The buffer memory is called a "recirculator," since the data in it are re-used many times. This technique is used in the VLA to implement spectroscopic correlation for up to 256 frequencies with only twice the number of multipliers needed for continuum.

Figure 4–8. A digital implementation of a simple correlator.

4. Spectroscopy

4.1. Why a spectroscopic correlator?

The observation of spectral line sources usually requires the analysis of the received bandwidth into a large number of sub-bands in order to measure the line profile and hence the radial velocity distribution of the emitting region. Even in the case of continuum sources, it is often desirable to have not just the cross power over the whole bandwidth, but the cross power spectrum. This is because even though the source has the same brightness distribution at all frequencies in the band, the spatial frequency (u-v coordinate) measured by each interferometer varies across the band. When the total bandwidth is large (for improved sensitivity) the final image is improved if account is taken of this, even if only one image of the integrated emission over the band is needed. For these reasons, correlators for most synthesis telescopes are required to measure the cross power spectrum on each baseline.

4.2. Design alternatives.

We have seen that there are two ways to build a correlator that measures the cross power in several sub-bands of the received bandwidth (Figures 4–4 and 4–5), according to whether the frequency analysis is done before or after the cross correlation. In the first case, the filter banks could be implemented in analog circuitry prior to digitization, or they could be implemented in digital circuitry afterwards. We thus have three major design types, all producing similar results. Intermediate or "hybrid" designs are also possible, where part of the frequency analysis is done with analog filters, part in pre-correlation digital filters, and part in a post-correlation discrete Fourier transform. The choice among these alternatives is primarily a matter of the cost of construction, which depends on the number of baselines, the number of sub-bands required, the total bandwidth, and the available technologies. However, there are also subtle differences in the results produced by the different types of correlators.

Let us develop an equation to describe the cost of a general correlator. Suppose that each analog filter costs C_a (independent of bandwidth, a fairly good approximation), a K–channel digital filter costs $C_d(K)$ per unit bandwidth (proportional to bandwidth, also a good approximation), and each cross-correlator costs C_x per unit bandwidth. Then a complete correlator for N antennas, total bandwidth $\Delta\nu$, and M sub-bands (frequency channels) will cost

$$C_{\mathrm{total}} = N[JC_a + C_d(K)\Delta\nu/(JL)] + \tfrac{1}{2}N(N-1)C_x\Delta\nu/(JK)\,, \qquad (4\text{–}13)$$

where J is the number of analog filter channels, K is the number of digital filter channels, and L is the number of post-correlation DFT channels. The cost can be minimized over J, K, and L subject to the constraint that the total number of channels is $M = JKL$. Notice that for sufficiently many antennas N the last term will dominate, and it is then advantageous to do as much processing before correlation as possible. The division between analog and digital processing depends on the relative costs of those technologies. A given type of digital circuit has a maximum speed of operation, but any bandwidth can be processed by having enough circuits operating in parallel; this is why the cost is assumed proportional to bandwidth. When the total bandwidth gets large (more than a few hundred MHz with present technologies), it becomes advantageous to use some analog filters.

From the above considerations alone, it seems that any large, spectroscopic correlator should be built with pre-correlation frequency analysis. This is indeed the trend for new designs. Such a correlator has become known as the "FX" type, after Chikada *et al.* (1984), meaning frequency analysis (F) before cross correlation (X); the opposite extreme, where the cross correlation function is computed first, is now called "XF." At this time, the only FX correlator in existence is at the Nobeyama Radio Observatory, although a larger one is being designed for the VLBA. For both types, the frequency analysis is usually done by via the FFT, so the "F" might also mean "Fourier." The FX architecture is not always the lowest cost, and its performance is subject to some complications that I'll discuss shortly. Besides, there are many XF correlators in use, including the VLA's correlator. You should therefore become familiar with both types.

4.3. The XF correlator.

The basic cross-correlation spectrometer is shown in Figure 4–3. Further details of a digital implementation are shown in Figure 4–9, where you should note several things: (1) the small delays $\delta\tau$ that determine the spacing of the correlation function measurements are implemented in a shift register, which means that they must be multiples of the sampling interval; (2) by including a self-multiplier for each digitized signal, the power in each signal can be determined (provided that at least 3-level quantization is used); (3) the quantization correction is applied to each cross-correlation measurement prior to computing the DFT.

Figure 4–9. A digital cross correlation spectrometer, with self-multipliers and quantization corrections.

The restriction of $\delta\tau$ to a multiple of the sampling period might seem to be no problem, because if the original signals have bandwidth $\Delta\nu$, then their cross power spectrum can have bandwidth at most $\Delta\nu$ also. Applying the sampling theorem to the cross power spectrum, we find that there is no loss of information if the cross correlation function is sampled at an interval $\delta\tau \leq 1/(2\Delta\nu)$, which is compatible with Nyquist sampling of the signals. But this restriction does lose information and lead to errors in the computed spectrum, for two reasons: First, as mentioned earlier, the *quantized* signals are not bandlimited to $\Delta\nu$, so neither is the cross power spectrum; sampling at only $2\Delta\nu$ causes the power outside $\Delta\nu$ to show up inside, a phenomenon called "aliasing." Second, the sampling theorem requires measurements at delays from $-\infty$ to $+\infty$, and the necessary truncation at a finite number of measurements usually has a significant effect. To see this, note that the sampling theorem allows us to write the cross power spectrum as an infinite sum

$$r(\nu) = \int_{-\infty}^{\infty} x(\tau)\, e^{-2\pi i\nu(\tau-\tau_0)}\, d\tau \tag{4–14a}$$

$$= \sum_{k=-\infty}^{\infty} x(\tau_0 + k\delta\tau)\, e^{-2\pi i\nu k\delta\tau}\, \delta\tau\,, \tag{4–14b}$$

where the second equation holds only within the bandwidth $0 \leq \nu \leq \Delta\nu <$

$1/(2\delta\tau)$. If the sum is truncated beyond $|k| = K$, the result may be written

$$\hat{r}(\nu) = \sum_{k=-\infty}^{\infty} \sqcap(k/2K)\, x(\tau_0 + k\delta\tau)\, e^{-2\pi i\nu k\delta\tau}\, \delta\tau \tag{4–15a}$$

$$= \int_{-\infty}^{\infty} \sqcap(\tau/2K\delta\tau)\, \text{Ш}(\tau/\delta\tau)\, x(\tau)\, e^{-2\pi i\nu(\tau-\tau_0)}\, d\tau \tag{4–15b}$$

$$= r(\nu) * \int_{-\infty}^{\infty} \sqcap(\tau/2K\delta\tau)\, \text{Ш}(\tau/\delta\tau)\, e^{-2\pi i\nu(\tau-\tau_0)}\, d\tau\,, \tag{4–15c}$$

where $\sqcap(\cdot)$ is the unit rectangle function and $\text{Ш}(\cdot)$ is the unit sampling function (Bracewell, 1978). The last integral may therefore be regarded as the bandpass function of a single channel; for large K, it is approximately $K\,\mathrm{sinc}(K\nu\delta\tau)$.

Equation 4–15 can be computed by the DFT, but only for certain values of ν. Normally, a length-K FFT is used, giving K complex results at $\nu = 0, 1/(K\delta\tau), \ldots, (K-1)/(K\delta\tau)$. Values at intermediate frequencies can be computed by means of a longer FFT with zeros inserted for $x(\tau_0 + k\delta\tau)$ when $|k| \geq K$.

The integrators normally accumulate data for a large number of samples before being read and reset to begin a new measurement. For example, if the sampling rate is 200 MHz (the Nyquist rate for 100 MHz bandwidth) and the integrating time is 1 sec, then 2×10^8 samples are accumulated. For this reason, the cost of the computing operations done after the integrators, including the FFTs, is generally negligible compared to earlier operations.

4.4. The FX correlator.

The architecture of a *digital* FX correlator is illustrated in Figure 4–10. The filter banks of Figure 4–4 have been replaced by discrete Fourier transforms that operate on successive segments of the digitized signals. In its simplest form, the correlator computes a length-K transform on non-overlapping segments of $2K$ samples each. Then the time available to compute one transform is $2K\Delta t$, the number of complex multiplies and adds required per transform is $K\log_2 K$, and one such transform is needed for each of N antennas in an array. Thus, the total computation rate for DFTs is $N\Delta\nu\log_2 K$ complex operations per second (COPS), assuming that $\Delta t = 1/(2\Delta\nu)$. A set of K complex cross multipliers and accumulators is needed for each of $N(N-1)/2$ baselines, but each must perform a multiply and add only every $2K$ samples, leading to a computation rate of $N(N-1)\Delta\nu/4$ COPS. So, altogether we need

$$c_{\mathrm{FX}} = \Delta\nu[N\log_2 K + N(N-1)/4] \tag{4–16}$$

COPS. The great attraction of this architecture for large systems is that c_{FX} increases only slowly with K and N. Compare this with the XF correlator, where

$$c_{\mathrm{XF}} = \Delta\nu K N(N-1)/2 \tag{4–17}$$

Figure 4–10. A digital FX spectrometer. In its simplest form, $K_1 = K_2 = K =$ number of frequency channels. To improve sensitivity, $K_2 = K/2$ (zero padding) and $K_1 > K_2$ (overlapping) are sometimes used, as described in the text.

if we count each real multiply-add as equivalent to 1/4 of a COPS.[3] To a reasonable approximation, the cost of one COPS is a constant of the technology; letting this cost be C_0, we see from comparing Equations 4–14 and 4–16 that $C_d(K) = C_0 \log_2 K$ and $C_x = C_0/2$ in this architecture.

Although the FX and XF correlators in Figures 4–9 and 4–10 can produce measurements of the cross power spectrum at the same K frequencies, they are not equivalent. This is because they produce different channel bandpass functions, as I'll now show. The DFT of a length-K segment of the (delayed) samples from antenna i is given by

$$V_i(\nu) = \sum_{k=-\infty}^{\infty} \sqcap(k/2K)\, v_i(k\delta t - \tau_0)\, \delta t\,, \tag{4–18}$$

for certain frequencies ν; and $V_j(\nu)$ is the similar result for the samples from antenna j (without the τ_0 delay). The estimated cross-power spectrum is the

[3] Actually, the arithmetic operations needed in the XF correlator are usually much simpler than in the FX because the latter must carry more bits of precision in order to avoid additional quantization noise; most operations in an XF correlator are on 1- or 2-bit samples. A proper comparison of the two architectures must take account of the resulting difference in cost per COPS.

average of the products of these transforms at each frequency:

$$\begin{aligned}\hat{r}(\nu) =& \langle V_i(\nu)\, V_j^*(\nu)\rangle \\ =& \sum_{k=-\infty}^{\infty} \sum_{k'=-\infty}^{\infty} \sqcap(k/2K)\sqcap(k'/2K)\, \langle v_i(\tau_0 + k\delta t) v_j(k'\delta t)\rangle\, e^{-2\pi i\nu(k-k')\delta t}\, \delta t \\ =& \sum_{k=-\infty}^{\infty} \sum_{k'=-\infty}^{\infty} \sqcap(k/2K)\sqcap(k'/2K)\, x_{ij}(\tau_0 + (k-k')\delta t)\, e^{-2\pi i\nu(k-k')\delta t}\, \delta t\, .\end{aligned} \tag{4–19}$$

The double sum may be rewritten by letting $l = k - k'$ as

$$\sum_{l=-\infty}^{\infty} W_l R_{ij}(\tau_0 + l\delta t) e^{-2\pi i\nu l\delta t} \delta t\, , \tag{4–20}$$

where

$$W_l = \begin{cases} |2K - l|, & \text{if } |l| < 2K \\ 0, & \text{otherwise.} \end{cases} \tag{4–21}$$

Using the unit sampling function $\text{Ш}(\cdot)$ to put this into integral form gives

$$\hat{r}(\nu) = r(\nu) * \int_{-\infty}^{\infty} 2K\Lambda(\tau/4K\delta t)\, \text{Ш}(\tau/\delta t)\, e^{-2\pi i\nu(\tau-\tau_0)}\, d\tau, \tag{4–22}$$

where $\Lambda(\cdot)$ is the unit triangle function. As before, the last integral is the bandpass function of each channel; for large K, it is approximately $2K\,\mathrm{sinc}^2(K\nu\delta t)$. This shows that the FX correlator response is equivalent to triangularly-weighted measurements of the cross correlation function over a range of $4K\delta t$ in delay. By comparison, the XF correlator response (Eq. 4–15) is a uniformly-weighted transform of direct measurements of the cross correlation over a range of $2K\delta t$. In both cases, the weighting can be modified by multiplying the data by an additional weighting function before the DFT, and this can be used to trade resolution for an improved shape of channel bandpass.

Some consequences of these results are:

(1) Equation 4–22 needs to be evaluated at intervals of no more than $1/(2K\delta t)$ in frequency in order to preserve all its information. This can be done with an FFT of (complex) length $2K$ by padding the input vector with zeros.

(2) The FX correlator suffers a reduction in sensitivity compared with the XF because samples from different segments are never cross correlated, even though the delay between them may be small. That is, the segmentation implies that not every opportunity to measure the cross correlation at a given delay is taken. This loss can be completely recovered if a new FFT is computed every sample time instead of every $2K$ sample times; i.e., the segments are made to overlap almost completely. However, this would eliminate

all of the cost advantage of the FX architecture. A compromise is to overlap partially, and it turns out that most of the sensitivity loss is recovered by a 50% overlap, where each sample is used in two successive segments.

Each of the above effects forces the FX correlator to cost about a factor of 2 more than Equation 4–16 indicated. Nevertheless, for large K and N it may have a significant advantage.

4.5. Quantization corrections.

The systematic effects of the quantization on the cross power spectrum can be eliminated in the XF correlator by applying the quantization correction to each cross correlation measurement prior to the DFT. Each measurement is then adjusted to what it would have been without quantization, except for the quantization noise. This is the arrangement is shown in Figure 4–9. If the digital cross correlation function is used without correction, then there will generally be a distortion of the spectrum whose form is hard to predict. Nevertheless, if the cross correlation function is small at all delays—that is, if the source is weak compared with the system noise—then the correction factor will be nearly the same at all delays, so the spectrum will be wrong only by a scale factor. For three level quantization, the departure from linearity exceeds 0.5% for correlation coefficients above about 0.3; for four level, this occurs above about 0.6. Notice that it is the correlation function of the whole bandwidth that matters, not each frequency channel; the source can be much stronger than the system noise in a few channels, and one might still be able to treat the quantization correction function as linear.

In the FX correlator, the quantization correction is more of a problem. Assuming that zero-padded transforms have been done, the estimated cross power spectrum will be available at sufficiently many frequencies that an inverse DFT will produce the (weighted) cross correlation function for delays $-2K\delta t$ to $2K\delta t$. The quantization correction can then be applied to each cross correlation, and the corrected function re-transformed to yield the corrected spectrum. While it may seem computationally wasteful to return to the delay domain, remember that these transforms are being done after integration, so they are far less frequent than those of the signal segments.

4.6. The Gibbs phenomenon.

Consider the situation illustrated in Figure 4–11, where the actual and computed cross power spectra are shown for signals from a unit-flux continuum source in the reference direction; this is the interferometer's gain vs. frequency function, and in the illustrated case the receivers have a fairly flat response. As you might expect, the computed spectrum shows ringing near the edges, where the true spectrum changes rapidly. This is the well-known "Gibbs phenomenon," which also occurs in autocorrelation spectrometers. The trouble is that if the computed spectrum from a continuum source is used as the complex gain for calibration purposes, then large errors can be made when a strong line source is observed. To see this mathematically, let $f(\nu)$ be the channel bandpass function given by the integral in Equation 4–15c; then the apparent complex gain on the i–j

baseline is $f(\nu) * [g_i(\nu)g_j^*(\nu)]$. This is what the correlator would measure for a unit-flux continuum source. When an unknown source whose true visibility is $V(\nu)$ is observed, the correlator measures

$$\hat{r}(\nu) = f(\nu) * [g_i(\nu)g_j^*(\nu)V(\nu)] \, . \tag{4–23}$$

Dividing by the apparent gain gives

$$\hat{V}(\nu) = \frac{f(\nu) * [g_i(\nu)g_j^*(\nu)V(\nu)]}{f(\nu) * [g_i(\nu)g_j^*(\nu)]} \, . \tag{4–24}$$

Notice that the convolution operations prevent cancelling of the gains, as might be desired. There are better ways of estimating $V(\nu)$ than simply taking the above ratio, such as deconvolving both the numerator and denominator, but these are not in common use. The situation can be somewhat improved by weighting the cross correlation before transforming, thereby smoothing the channel bandpass, but this sacrifices some frequency resolution. In practice, measurements near the band edges must be discarded for a variety of reasons, of which the Gibbs phenomenon is only one.

Because the cross power is complex, the Gibbs phenomenon behaves somewhat differently here than in the autocorrelation spectrometer. Both $r(\nu)$ and $\hat{r}(\nu)$ are Fourier transforms of real functions, so they are Hermitian: $r(\nu) = r^*(-\nu)$. If the passband extends to near zero frequency, as it usually does at the input to a digital correlator, then the imaginary part of the gain makes a sharp change at $\nu = 0$, whereas the real part does not. This means that not even observations of a continuum source will be correctly calibrated by using the computed cross spectrum, unless the cross powers of the source and calibrator happen to have the same phase.

5. Delay Resolution and Fringe Rotation Effects

In the foregoing discussion, I regarded the correlator as being responsible for estimating the cross power spectrum of whatever two signals are presented to it. This tacitly assumes that the correlation is not changing too rapidly; it must be reasonably constant during the time required to complete a measurement. As was shown in Lecture 2, this can be achieved by including in one signal path a variable instrumental delay that is continuously adjusted to compensate for the rapid change of geometric path delay caused by Earth rotation. Indeed, we have now seen that it is convenient to implement this delay after digitization, and to consider it part of the correlator. Lecture 2 also pointed out that implementation of the delay after frequency conversion(s) requires that a compensating phase shift also be added to the net local oscillator signal, or else the correlation function will be phase modulated by $\nu_{\mathrm{LO}}\tau_g$. This phase shift is called "fringe rotation."

In practice, the delay cannot be set to arbitrary precision, and receivers are sometimes built without including fringe rotation in their local oscillators. In this section, I shall evaluate the effects of these limitations.

Figure 4–11. **(a)** The cross power spectrum resulting from a continuum source of unit flux in the reference direction: "true complex gain." Note the nonzero phase. **(b)** The computed cross power spectrum with 16 delays.

Assume that the required delay can be calculated accurately and to arbitrary precision, but can only be set to discrete values spaced by $\Delta\tau$. This means that there will be a delay error, or difference between the required value and the setting, which can be kept between $-\Delta\tau/2$ and $+\Delta\tau/2$. During the integrating time of the correlator, variation of the geometrical delay may cause the delay setting to change by many steps (and thus cause the delay error to pass through its range many times), or it may be slow enough so that the delay setting stays constant (and the error changes only slightly). Typical modern telescopes experience both extremes in different parts of the sky and on different baselines. The effect of this delay error depends on which case occurs, and on the bandwidth of the signals.

To evaluate the effect, we may simply average the cross correlation over the correlator integrating time, including the time-varying delay error. Letting $\tau = \tau_0 + \delta\tau$, where $\delta\tau$ is the delay error, Equation 4–5 gives

$$R_{ij}(\tau_0 + \delta\tau) = 2\int_0^\infty r_{ij}(\nu) e^{2\pi i\nu\delta\tau}\, d\nu\,. \tag{4–25}$$

Now if the signals have rectangular spectra, then $r_{ij}(\nu)$ is constant with frequency up to the bandwidth $\Delta\nu$, so

$$R_{ij}(\tau_0 + \delta\tau) = R_{ij}(\tau_0)\frac{1}{\Delta\nu}\int_0^{\Delta\nu} e^{2\pi i\nu\delta\tau}\, d\nu\,. \tag{4–26}$$

If $\delta\tau$ is constant during the integrating time, this shows that the complex cross correlation is reduced by a complex factor. If $\delta\tau$ varies, then the result must be averaged over the variation. For a spectroscopic correlator, replace $x(\tau)$ in Equation 4–4 with $x(\tau + \delta\tau)$ and apply the shift theorem of Fourier transforms, obtaining

$$\hat{r}_{ij}(\nu) = r_{ij}(\nu) e^{2\pi i \nu \delta\tau} \,. \tag{4–27}$$

This shows that there is a phase shift proportional to frequency and to delay error. The effect is slightly modified for a practical DFT correlator (as in Fig. 4–5) because of the finite length of the transform, but Equation 4–27 holds fairly well for practical numbers of points.

Equations 4–26 and 4–27 are evaluated in Table 4–2 for some situations of practical interest. Two sizes of $\Delta\tau$ are considered: half the reciprocal bandwidth (one sample time at the Nyquist rate), and one-sixteenth as much. In the fast delay case (τ_0 changing by many $\Delta\tau$ per integration), the loss in amplitude for the continuum and for the highest spectrometer frequency (worst channel) are given. In the slow delay case (τ_0 nearly constant), the continuum amplitude loss is also given. In all cases there is also a phase shift. Since these effects are all calculable, appropriate corrections can be applied to the data; but the amplitude losses represent an irrecoverable drop in sensitivity, since there is no corresponding reduction in noise.

Table 4–2. Delay Error Loss Factors

	— Delay Resolution^{-1} —	
	Nyquist	16 × Nyquist
1. Fast Delay		
A. Continuum	0.9664	0.99987
B. Band Edge	0.9003	0.9996
2. Slow Delay		
A. Continuum, worst case	0.9003	0.9996

Table 4–2 shows that if the delay resolution is a small fraction of the reciprocal bandwidth, then the losses can be kept very small. But if all of the delay is implemented after sampling, then no finer delay resolution than one sample time can be achieved; so Nyquist sampling might be thought to force acceptance of the larger losses in the table. One solution is to build samplers whose sampling phase can be adjusted on a fine scale; this has been done in the VLA, but in some situations this may not be practical. For example, in VLBI the sampling must be done during observing, but the correlation will not be done until much later. At observe time, the source position and clock settings may not be known to sufficient accuracy to determine the optimum sampling phase. At correlate time, this information is available but the delay can now be set only to within one sample time.

Next, consider the case where the delay is implemented after conversion to a low frequency (e.g., baseband, for digital delays), but no compensating phase shift (fringe rotation) is applied to the local oscillator. It can be shown by a straightforward extension of the results of Lecture 2 that a simple correlator

Figure 4–12. A simple correlator with fringe rotation.

(like Fig. 4–1) produces the output

$$x'(\tau_g) = x(\tau_g)\cos 2\pi\nu_{\rm LO}\tau_g + \tilde{x}(\tau_g)\sin 2\pi\nu_{\rm LO}\tau_g\,, \qquad (4\text{–}28)$$

where $x(\tau)$ is the correlation function of the signals at the antennas, which is what would be measured by the correlator if fringe rotation were included; $\nu_{\rm LO}$ is the net local oscillator frequency; and the delay is set to $\tau_0 = \tau_g$. Note that τ_g is changing with time, perhaps rapidly, due to Earth rotation; so this result applies only if the correlator averaging time is short enough. To obtain a direct estimate of $x(\tau_g)$, and to allow use of longer averaging times, the technique of Figure 4–12 can be used. Here the correlator is modified by multiplying the cross product by an appropriate quasi-sinusoid prior to averaging. This sinusoid is called a "fringe function"; now the job of fringe rotation has been moved from the local oscillator to the correlator.

If the averaging time is an integral number of cycles of the fringe function, then the correlator of Figure 4–12 produces an unbiased estimate of $x(\tau_g)$. Nevertheless, this method has some disadvantages. One must integrate for at least one "fringe," and sometimes $\nu_{\rm LO}\tau_g$ changes too slowly for this. More importantly, the signal-to-noise ratio obtained is worse than that of the Figure 4–1 correlator (with fringe rotation applied to the LO, if any) by $\sqrt{2}$. This is because the fringe function is near zero much of the time (a detailed derivation is left as an exercise). One way to overcome this is to use the more complicated fringe rotator of Figure 4–13. Here both cosine and sine fringe rotators are used, and the results are combined with a $\pi/2$ phase shift before integrating. This makes use of the second term in Equation 4–28, and gives a signal-to-noise ratio equal to that of Figure 4–1. Such an arrangement is feasible even if the signals are digitized (since a digital implementation of the $\pi/2$ phase shift, or Hilbert transform, is possible), but to my knowledge it has not yet been used in radio astronomy.

There is another way to recover the full signal-to-noise ratio that would have been obtained with LO fringe rotation, but it applies only to spectroscopic correlators. The method is illustrated for an XF correlator in Figure 4–14. Note that the order of the fringe rotation and cross correlation multiplications has

Figure 4–13. A simple correlator with "single-sideband" fringe rotation.

been interchanged compared with Figure 4–13, but that this has no effect since multiplication is associative; thus only one fringe rotator is needed for all delays. If we were to use a sine fringe function rather than the cosine, it would have two effects on the results: after the DFT, the expected value of the (complex) result at each frequency would change phase by $\pi/2$ (i.e., real and imaginary parts would be interchanged); and the noise would be different. In fact, it can be shown that the noises in the two cases would be independent. Therefore, if the spectrum is obtained *both* ways (sine and cosine fringe rotation) and the results are averaged (after correcting for the phase difference), the signal-to-noise ratio is improved by $\sqrt{2}$. (Again, the proof is left as an exercise. You will probably find it easier to do after studying Lecture 7.) This method is quite expensive, since it doubles the required size of the correlator; the correlation at all delays must be measured simultaneously for both sine and cosine fringe functions.

An approach similar to that of Figure 4–14 can be taken with an FX correlator. The outputs of the sine and cosine fringe rotators are each segmented and transformed in an FFT, doubling the required rate of FFT computation. However, the averaging of the results can now be done *before* the cross multipliers, so that part of the correlator need not be doubled. This is a significant advantage when the number of antennas is large, because then the cross multiplication dominates the correlator cost. Furthermore, it turns out that the required averaging of the two FFT results can be done automatically by using a single, double-length FFT and discarding half of the output points.

In VLBI, most receivers have been implemented without fringe rotation in the LO, and the double-size spectroscopic correlator method has been extensively used to obtain the best signal-to-noise ratio, even for continuum observations. This has made sense because, until recently, most correlators have been small, handling typically 3 to 5 antennas at once, and with a relatively small bandwidth. In such a situation, slow and inexpensive digital electronics can be used, and not much of it is needed; the cost is dominated by other components, such as tape recorders. Also, it is inconvenient to install LO fringe rotation at many VLBI stations that were originally designed for single-dish work. But with new systems now being designed, the bandwidths and numbers of antennas are larger,

Figure 4–14. A spectroscopic cross correlator with fringe rotation.

vastly increasing the correlator size and making LO fringe rotation a reasonable option.

References

Bracewell, R. N. (1978), *The Fourier Transform and Its Applications*, Second Edition, Chapters 4 and 5, McGraw–Hill, New York.

Chikada, Y., Ishiguro, M., Hirabayashi, H., Morimoto, M., Morita, K.-I., Miyazawa, K., Nagane, K., Murata, K., Tojo, A., Inoue, S., Kanzawa, T., and Iwashita, H. (1984), "A digital FFT spectro-correlator for radio astronomy", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 387–404.

Cooper, B. F. C. (1970), "Correlators with two-bit quantization", *Aust. J. Phys.*, **23**, 521–527.

D'Addario, L. R. (1984), "Minimizing storage requirements for quantized noise", VLBA Memorandum No. 332, NRAO.

D'Addario, L. R., Thompson, A. R., Schwab, F. R., and Granlund, J. (1984), "Complex cross correlators with three level quantization: design tolerances", *Radio Science*, **19**, 931–945.

Schwab, F. R. (1979), "Quantization correction of VLA correlation measurements", VLA Computer Memorandum No. 150, NRAO.

Shannon, C. E. (1949), "Communication in the presence of noise", *Proc. IRE*, **37**, 10–21.

Van Vleck, J. H. and Middleton, D. (1966), "The spectrum of clipped noise", *Proc. IEEE*, **54**, 2–19.

5. CALIBRATION AND EDITING

EDWARD B. FOMALONT AND RICHARD A. PERLEY

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 14 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. The general organization of the lecture is:

1. Introduction
2. Basic Considerations
3. Initial Calibrations

4. Routine Corrections
5. Bandpass Calibration
6. Editing
7. Final Calibration Using Radio Sources
8. Application of the Calibration Values
9. Polarization Calibration
10. Additional Topics.

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

[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.

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

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

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:

$$\begin{aligned}\Delta\phi_g = 2\pi\nu\Delta\tau_g = \frac{2\pi}{\lambda}\big(&\Delta L_x \cos H \cos\delta - \Delta L_y \sin H \cos\delta + \Delta L_z \sin\delta \\ &+\Delta\alpha \cos\delta(L_x \sin H + L_y \cos H) \\ &+\Delta\delta(-L_x \cos H \sin\delta + L_y \sin H \sin\delta + L_z \cos\delta)\big)\,,\end{aligned} \tag{5–3b}$$

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 5–3b 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

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

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 $\Delta t_{\rm int} \sim 2D_{\rm (m)}/B_{\rm (km)}$, where $B_{\rm (km)}$ is the length of the baseline in km and $D_{\rm (m)}$ 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 9 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. Most arrays shut down operation when the wind becomes excessive.

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 13. 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 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 16 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 time-keeping 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 and 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 $P_{\rm tot}$, water-vapor partial pressure P_v, and temperature T, the excess propagation path can be roughly estimated. An approximate formula is

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

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. 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 and 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). 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 $T_{\mathrm{atm}}(1 - e^{-\tau_a \sec z})$, where τ_a is the zenith opacity of the atmosphere and T_{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

Freq.	α_0	α_1
15 GHz	.013	.0002
22 GHz	.026	.015
35 GHz	.039	.0039
90 GHz	.039	.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 $k^{\rm th}$ 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 if 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 16), 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 grey-scale 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. 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.

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.

For convenience, the complete calibration Equation 5–4 is reproduced here,

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

where

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

In this section we will investigate how to fully exploit this relationship.

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. 7–3 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 16.

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 9. 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 the VLBI lectures 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.

The strategy of picking the correct calibrator for the observations is discussed in Section 7 of Lecture 24. 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.

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 $L \star R$. 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 and 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.

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 15), 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, 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 Section 2 of Lecture 13. 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 and 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. and Napier, P. J. (1978), "Preliminary polarization calibration of the VLA at 6 cm", VLA Test Memorandum No. 125, NRAO.

Brosche, P., Wade, C. M., and Hjellming, R. M. (1973), "Precise positions of radio sources. IV. Improved solutions and error analysis for 59 sources", *Astrophys. J.*, **183**, 805–818.

Carter, W. E., Robertson, D. S., and MacKay, J. R. (1985), "Geodetic radio interferometric surveying: application and results." *J. Geophys. Res.*, **90**, 4577–4587.

Clark, B. G. (1973a), "VLA telescope pointing analysis", VLA Computer Memorandum No. 104, NRAO.

Clark, B. G. (1973b), "Ephemeris routines for the VLA—specification considerations", VLA Computer Memorandum No. 105, NRAO.

Clark, B. G. (1982a), "The new IAU precession procedures", VLA Scientific Memorandum No. 145, NRAO.

Dewdney, P. (1987), "Investigating the use of tiltmeters to correct VLA antenna pointing—Part I. Initial measurements and analysis", VLA Test Memorandum No. 148, NRAO.

Fomalont, E. B. and Sramek, R. A. (1977), "The deflection of radio waves by the Sun", *Comments on Astrophys.*, **7**, 19–33.

Sramek, R. A. (1983), "VLA phase stability at 22 GHz on baselines of 100 m to 3 km", VLA Test Memorandum No. 143, NRAO.

Thompson, A. R. (1980), "Closure accuracy and related instrumental tolerances", VLA Electronics Memorandum No. 192, NRAO.

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

Uson, J. (1986), "Atmospheric opacity at the VLA", VLA Scientific Memorandum No. 157, NRAO.

Wade, C. M. (1970), "Precise positions of radio sources. I. Radio measurements", *Astrophys. J.*, **162**, 381–390.

6. IMAGING

Richard A. Sramek and Frederic R. Schwab

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),

$$\mathcal{A}(l,m)I(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{6–1}$$

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 (say, $\mathrm{W\,m^{-2}\,Hz^{-1}}$), 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 arc second is used instead. The units are determined by the normalization of Equation 6–1.

Equation 6–1 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 small, model fitting is feasible—and sometimes useful (see Lecture 11). 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 6–1, 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)

$$I^D(l,m) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u,v)V'(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{6–2}$$

[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) present a problem because the primary beam patterns rotate on the sky, as functions of parallactic angle. See Lecture 5.

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 6–2. The first, called the 'direct Fourier transform' method,[2] approximates $I^D(l,m)$ by brute-force evaluation of the sum

$$\frac{1}{M}\sum_{k=1}^{M} V'(u_k,v_k)e^{2\pi i(u_k l+v_k m)}\,. \tag{6–3}$$

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 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] 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 invertible; usually, of course, the term 'transform' is reserved for invertible transformations.

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',

$$S(u,v) = \sum_{k=1}^{M} \delta(u-u_k, v-v_k)\,. \tag{6–4}$$

It is useful to introduce a second generalized function, called the *sampled visibility function* or, alternatively, the *u-v measurement distribution*,[3]

$$V^S(u,v) \equiv \sum_{k=1}^{M} \delta(u-u_k, v-v_k) V'(u_k,v_k)\,. \tag{6–5}$$

That is, $V^S = SV'$. Let $\mathbf{F}$ denote the Fourier transform operator. Equation 6–2 can be rewritten

$$I^D = \mathbf{F}V^S = \mathbf{F}(SV')\,. \tag{6–6}$$

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),

$$I^D = \mathbf{F}S * \mathbf{F}V'\,, \tag{6–7}$$

where $*$ denotes convolution. For a point source of unit strength, centered at position (l_0, m_0), $|V'(u,v)| \equiv 1$, and $\mathbf{F}V'$ is the (shifted) Dirac δ-function: $\delta(l-l_0, m-m_0)$. So the point source response of the array, i.e., the *synthesized beam*, is given by $B = \mathbf{F}S * \delta = \mathbf{F}S$. Equation 6–7 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 Expression 6–3, is *exactly* I^D. That is to say, that—assuming δ-function sampling—$I^D(l,m)$, as defined by Equation 6–2, is given exactly by a discrete summation, Expression 6–3, and that Equation 6–7 holds 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.

[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 12, 13, and 24.

2.2. Weighting functions for control of the beam shape.

In analogy to Equation 6–4, a *weighted sampling function*, or *weighted sampling distribution*, can be written as

$$W(u,v)=\sum_{k=1}^{M} R_k T_k D_k \delta(u-u_k, v-v_k)\,. \tag{6–8}$$

And, in analogy to Equation 6–5, one can define a *weighted, sampled visibility function*, or *weighted and sampled measurement distribution*, V^W according to $V^W = WV'$, or, explicitly,

$$V^W(u,v)=\sum_{k=1}^{M} R_k T_k D_k \delta(u-u_k, v-v_k) V'(u_k, v_k)\,. \tag{6–9}$$

The coefficients R_k, T_k, and D_k are weights assigned the visibility points. These data points may represent time-averages of visibility measurements spaced along the loci of the interferometer u-v tracks. R_k is a weight that indicates the reliability of the k^{th} visibility datum. It may depend on the amount of integration time, the system temperature, and the bandwidth used for that data point. There is no control of R_k in the image formation, so no further mention is made of it here.

The density weight D_k and the taper T_k can be specified in many Fourier transform imaging programs, to 'fine-tune' the beam shape. 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 many arrays more data points fall in the inner region of the u-v plane than fall further out. This tends to give higher weight to the low spatial frequencies. The natural sampling may impair effective deconvolution or mask interesting features of I.

The D_k and the T_k are used to control, to some extent, the beam shape. The T_k are used to weight down 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.

2.2.1. 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. 6–1).

Figure 6–1. A Gaussian u-v taper with dispersion $\sigma = 1$ km.

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) is $\theta_{\mathrm{HPBW}} = 0.37/\sigma$ with θ in radians and σ in wavelengths. Translated into common units, $\theta_{\mathrm{HPBW}} = 0.77\lambda_{(\mathrm{cm})}/\sigma_{(\mathrm{km})}$ arc-seconds. 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 $\theta_{\mathrm{HPBW}} = 2.0\lambda_{(\mathrm{cm})}/a_{(\mathrm{km})}$ arc-seconds, where a is the radius of the aperture. Real-life observational geometries and u-v coverages often produce larger θ_{HPBW} and, frequently, elongated beams. Examples of the VLA point source response with different u-v tapers are shown in Figure 6–2.

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 under-sampled 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.

Figure 6–2. 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.

2.2.2. 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:

$$\begin{aligned} D_k &= 1\,, \qquad \text{called } \textit{natural weighting}, \\ \text{and}\quad D_k &= \frac{1}{N_s(k)}\,, \qquad \text{called } \textit{uniform weighting}, \end{aligned}$$

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.) In many Fourier transform imaging programs s is a free parameter selected by the user.

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.

Sometimes, especially in the VLA 'snapshot' mode of observing (see Appendix A7 of Lecture 24), 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 6–3.

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 the observed values.[5] There are many ways to achieve this interpolation (see, e.g., Thompson and 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

[4] In the AIPS implementation, these blocks are called 'uniform weight boxes'.

[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.

Figure 6–3. The effect of different weighting functions on a VLA 'snapshot' image of a point source.

the 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

$$\sum_{k=1}^{M} C(u_c - u_k, v_c - v_k) V^W(u_k, v_k) \,. \tag{6–10}$$

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 Expression 6–10 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

$$V^R = R\left(C * V^W\right) = R\left(C * (WV')\right), \qquad (6\text{–}11)$$

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

$$R(u,v) = \text{Ш}(u/\Delta u, v/\Delta v) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \delta(j - u/\Delta u, k - v/\Delta v)\,. \qquad (6\text{–}12)$$

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 $\mathbf{F}V^R$ can be obtained by a discrete Fourier transform. Thus $\mathbf{F}V^R$ can be calculated by the FFT algorithm.

$\mathbf{F}V^R$—after normalization, and after one simple correction—is what you have been seeking: a 'dirty' image—a cheap approximation to I^D. Denote $\mathbf{F}V^R$ by $\widetilde{I}^D$.

Applying the convolution theorem to Equation 6–11, $\widetilde{I}^D$ is given by

$$\widetilde{I}^D = \mathbf{F}R * \left[(\mathbf{F}C)\left(\mathbf{F}V^W\right)\right] = \mathbf{F}R * \left[(\mathbf{F}C)\left(\mathbf{F}W * \mathbf{F}V'\right)\right]. \qquad (6\text{–}13)$$

(Please refer now to Fig. 6–4 for a graphical interpretation of Eq. 6–13 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),

$$(\mathbf{F}R)(l,m) = \Delta u\,\Delta v\,\text{Ш}(l\Delta u, m\Delta v) = \Delta u\,\Delta v \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \delta(j - l\Delta u, k - m\Delta v)\,. \qquad (6\text{–}14)$$

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 $\mathsf{F}R$ (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., $c \equiv \mathsf{F}C$). 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

$$\widetilde{I}^D_c(l,m) = \frac{\widetilde{I}^D(l,m)}{c(l,m)}\,, \tag{6–15}$$

and the 'grid-corrected' beam by

$$\widetilde{B}^D_c(l,m) = \frac{\widetilde{B}^D(l,m)}{c(l,m)}\,. \tag{6–16}$$

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 6–11 and $\mathsf{F}R$ in Equation 6–13. 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]

[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 and 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').

Finally, $\widetilde{I}_c^D$ and $\widetilde{B}_c^D$ both are normalized by a scaling factor selected so that the peak of $\widetilde{B}_c^D$ is of unit flux density. One may as well not alter the notation to reflect this, since it is a trivial operation.

Figure 6–4 *(pp. 128–129)***.** 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 6–4. 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 6–5. 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 reciprocal 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. The result is shown in panel **(o)**. This is the end product, the 'dirty image' that is supplied to the deconvolution programs.

Figure 6–4 *(Caption is on p. 127)*. *(Continued on next page.)*

Figure 6–4 *(Continued).*

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 $\widetilde{I}_c^D$ is a good approximation to I^D—that is, Equation 6–13 becomes

$$\widetilde{I}_c^D = \mathsf{F}W * \mathsf{F}V' , \tag{6–17}$$

—and then the usual convolution relation between I^D, B, and I is approximately valid with $\widetilde{I}_c^D$ and $\widetilde{B}_c^D$ substituted for I^D and B, respectively. Note, however, 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.

3.2. Aliasing.

Due to the presence of $\mathsf{F}R$ in Equation 6–13 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 (Sec. 4.6).

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.

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:

(1) a 'pillbox' function,
(2) a truncated exponential,
(3) a truncated sinc function ($\operatorname{sinc} x \equiv \frac{\sin \pi x}{\pi x}$),
(4) an exponential multiplied by a truncated sinc function, and
(5) 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, 1980a):

(1) *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.

(2) *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.

(3) *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.

(4) *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.

(5) *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* con-

[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 6–5. The response to a source, as a function of distance from the image center, for two typical u-v convolving functions.

centration 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 (1984a) for further discussion and additional references.

Figure 6–5 shows the Fourier transforms of two typical gridding convolution

Figure 6–6. 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 $\psi_{-\frac{1}{2},0}$, $m = 6$. Adapted from Schwab (1984a).

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 $\exp \times \mathrm{sinc}$ 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 6–6 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 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.) Schwab (1978) and Greisen (1979) show plots of q for these convolving functions and for many others.

The pillbox, exponential, and sinc functions do not give as effective alias rejection as the $\exp \times \mathrm{sinc}$ or the spheroidal. The $\exp \times \mathrm{sinc}$ has somewhat smaller

Figure 6–7. The effects of aliasing: **(a)** a point source at the field center; **(b)** the same source near the image edge; **(c)** the source below the lower image edge appears as an aliased image at the upper image edge, with pillbox convolution; **(d)** with $\exp \times \mathrm{sinc}$ convolution, the aliased source is greatly attenuated, but the sidelobe response remains the same.

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 1984a).

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. 6–7). 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 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 $(\mathsf{F}g)(\mathbf{x}) = \alpha^{-n}(\mathsf{F}f)(\mathbf{x}/\alpha)$. 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—the reason that 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.

Most Fourier transform imaging programs do not work quite as described above. Often the tapering, introduced in Equation 6–8, 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 $\mathsf{F}T$, which, if inserted into Equation 6–13, 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 $\mathsf{F}R$, but one with a very long period, $100/\Delta u$. The grid correction given by Equation 6–15 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 (1984a) gives cheap and accurate rational approximations to the spheroidal functions; the step function approximation is unnecessary.)

4.3. Non-coplanar baselines.

In Equation 6–1 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 6–1 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

$$V(u, v, w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\mathcal{A}(l, m)I(l, m)}{\sqrt{1 - l^2 - m^2}} e^{-2\pi i\left(ul+vm+w\left(\sqrt{1-l^2-m^2}-1\right)\right)}\, dl\, dm\,. \tag{6–18}$$

This can be rewritten as

$$V(u, v, w)e^{-2\pi i w} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\mathcal{A}(l, m)I(l, m)}{\sqrt{1 - l^2 - m^2}}\, \delta(n - \sqrt{1 - l^2 - m^2})e^{-2\pi i(ul+vm+wn)}\, dl\, dm\, dn\,. \tag{6–19}$$

Now, by sampling V, weighting by $e^{-2\pi i w}$ and by the Fourier kernel, and integrating over (u, v, w), one obtains an analog of Equation 6–2,

$$I^{3D}(l, m, n) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u, v, w)V(u, v, w)e^{-2\pi i w}e^{2\pi i(ul+vm+wn)}\, du\, dv\, dw\,, \tag{6–20}$$

which (*cf.* Eq. 6–19) is equal to a three-dimensional convolution—the convolution of

$$I^3(l, m, n) \equiv \frac{\mathcal{A}(l, m)I(l, m)}{\sqrt{1 - l^2 - m^2}}\, \delta(n - \sqrt{1 - l^2 - m^2})\,, \tag{6–21}$$

with

$$B^{3D}(l,m,n) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u,v,w)e^{2\pi i(ul+vm+wn)}\,du\,dv\,dw\,. \qquad (6\text{–}22)$$

Note that I^3 is a distribution confined to the celestial sphere and that B^{3D} 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 6–20. In applying the 'direct Fourier transform' method, one simply uses a discrete summation, in analog to Equation 6–3. In the FFT method, w-terms need to be inserted into Expression 6–10, 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. At the VLA, such a 3–D imaging capability was designed into the 'pipeline' imaging system, but it has seldom been used. Rick Perley, in Lecture 14, will report on more recent efforts to deal in software with the non-coplanar baseline problem.

5. 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, $\widetilde{I}_c^D$. 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 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, too, in spectral line observing, primarily because spectral line data reduction is computationally very expensive, and because narrow bandwidths lead to low signal-to-noise ratios.

In wide-field imaging, deconvolution is the real problem in trying to cope with non-coplanar baselines. Because simple 2–D deconvolution itself is an extremely expensive operation, there has been little progress to date in obtaining high quality (deconvolved) images taking proper account of sky curvature and non-coplanar baselines, via any sort of three-dimensional deconvolution technique (to complement the 3–D imaging techniques described in Sec. 4.3). Data storage is another problem. Typically, non-coplanar baseline

[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 $e^{2\pi iw}$, which cancels the multiplication by $e^{-2\pi iw}$ in Equation 6–20.

effects are an important concern in the largest images; but computer storage is often barely adequate for the number of points, or 'pixels', required in just the l- and m-coordinates. A crude approach which has yielded some useful results involves *mosaicing*—constructing 'patchwork' images, each piece computed with the '$w(n-1)$'-correction appropriate to the center of the patch. This approach, which is used in the AIPS program 'MX', combining linear imaging with deconvolution, is described in Lecture 8. Because sidelobes from a source in any one patch fall into each of the other patches of the mosaic, the deconvolution operation must work in parallel on the patches. This necessitates repeated re-gridding of data.

More recent developments in 3–D imaging and 3–D deconvolution will be described in Lecture 14 by Rick Perley.

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 Figure 6–4.

References

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

Dym, H. and McKean, H. P. (1972), *Fourier Series and Integrals*, Academic Press, New York.

Erickson, W. C., Mahoney, M. J., and Erb, K. (1982), "The Clark Lake Teepee-Tee telescope", *Astrophys. J. Suppl. Ser.*, **50**, 403–420.

Greisen, E. W. (1979), "The effects of various convolving functions on aliasing and relative signal-to-noise ratios", VLA Scientific Memorandum No. 131, NRAO.

Schwab, F. R. (1978), "Suppression of aliasing by convolutional gridding schemes", VLA Scientific Memorandum No. 129, NRAO.

Schwab, F. R. (1980a), "Optimal gridding", VLA Scientific Memorandum No. 132, NRAO.

Schwab, F. R. (1984a), "Optimal gridding of visibility data in radio interferometry", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 333–346.

Stratton, J. A. (1935), "Spheroidal functions", *Proc. Nat. Acad. Sci. U.S.A.*, **21**, 51–56.

Thompson, A. R. and Bracewell, R. N. (1974), "Interpolation and Fourier transformation of fringe visibilities", *Astron. J.*, **79**, 11–24.

7. SENSITIVITY

PATRICK C. CRANE AND PETER J. NAPIER

1. INTRODUCTION

In this lecture we analyze the sensitivity of a synthesis array, derive general expressions for r.m.s. noise levels and evaluate these expressions for the particular case of the VLA. It is important to note that we will consider only the noise effects of the observed radio source itself and of additive random noise. By additive noise we mean white, Gaussian noise that is added to the astronomical signal received by an antenna before cross correlation with the output from another antenna. The sources of the additive noise are the 3 K microwave background, the galactic background, thermal noise generated by atmospheric emission, thermal noise picked up from the ground, thermal noise due to attenuation in the input microwave feed and waveguide structure, noise from the injected calibration signal and noise generated in the low-noise receiver itself. Just as the sensitivity of a single-antenna radio telescope is often not limited by random noise but is determined, rather, by effects such as confusion and gain instability, there are many effects other than random noise which limit the sensitivity of a synthesis array. The most important of these effects, which are *not* considered here, include errors in calibrating the complex gain of the instrument, atmospheric amplitude and phase instabilities, effects of sidelobes and confusing sources, radio-frequency interference, DC offsets in the correlators, and the distortions caused by a non-negligible bandwidth. Some of these effects introduce artifacts (e.g., stripes) into the image, while others mimic additive random noise by merely raising the noise level in the image.

2. DEFINITION OF SYSTEM TEMPERATURE

Figure 7–1 shows a schematic diagram for a two-element, single-multiplier, correlation interferometer. All of the electronics from the output of the feed horn to the input to the multiplier are represented by a single receiver with power gain G and bandwidth $\Delta\nu$ at the multiplier input. In a practical interferometer the signal may undergo many frequency conversions between the feed and the multiplier input, but G and $\Delta\nu$ are still well-defined quantities.

For the purpose of analyzing the sensitivity of this simple interferometer, it is useful to replace the antenna at the input to the receiver with a matched termination having physical temperature T. The power P entering the receiver from this termination is given by[1]

$$P = k_B T \Delta\nu \,, \tag{7–1}$$

where k_B is Boltzmann's constant (1.38×10^{-23} joule K^{-1}).

Suppose that the antenna is pointed at a piece of blank sky that contains no astronomical radio sources (other than the 3 K microwave background and the galactic background). In this case all of the power at either input to the correlator is due to additive noise resulting from phenomena listed in Section 1. Call this power at the correlator input P_N. Then we define the system temperature, $T_{\rm sys}$, to be the physical temperature of a matched termination on the input of the receiver (now assumed to be noiseless) which will produce P_N at the correlator input. That is

$$k_B T_{\rm sys} \Delta\nu \, G = P_N \,. \tag{7–2}$$

Suppose that the antenna now points at a radio source. If G remains constant, the power at the correlator input will increase to $P_N + P_a$, where P_a is the additional power collected by the antenna from the radio source. We define the antenna temperature due to the source, T_a, to be the increase in temperature of a termination on the receiver input needed to produce an increase in power of P_a at the correlator. That is

$$k_B T_a \Delta\nu \, G = P_a \,. \tag{7–3}$$

Note that we have defined $T_{\rm sys}$ to include only the effects of additive noise. Sometimes $T_{\rm sys}$ is defined to include the noise due to the radio source under observation, but it will be convenient in the following analysis to separate T_a from $T_{\rm sys}$. In most practical situations this point is not important because $T_a \ll T_{\rm sys}$. It is interesting to break $T_{\rm sys}$ up into its component parts:

$$T_{\rm sys} = T_{\rm bg} + T_{\rm sky} + T_{\rm spill} + T_{\rm loss} + T_{\rm cal} + T_{\rm rx} \,, \tag{7–4}$$

where

$T_{\rm sys}$ = Total system temperature excluding noise contribution from the observed radio source;
$T_{\rm bg}$ = Noise contribution from microwave and galactic backgrounds;
$T_{\rm sky}$ = Noise contributed from atmospheric emission;
$T_{\rm spill}$ = Noise contribution due to ground radiation scattering into the feed from the subreflector edge, feed legs, dish edge, etc.;

[1] Remember that this is an approximation. It is equivalent to the Rayleigh–Jeans approximation to Planck's black-body radiation law, and holds when $h\nu \ll k_B T$, where ν is the frequency and h is Planck's constant (6.63×10^{-34} joule sec). The approximation is valid provided that the frequency is not too high and the temperature is not too low. It is in error by 4% in estimating the noise power available from a 22 GHz termination at 15 K. For some of the new millimeter-wavelength interferometers it may be necessary to use the correct Planck equation when analyzing and measuring sensitivities. See, for example, Kraus (1986).

Figure 7–1. Simple block diagram of a two-antenna, single-multiplier, correlation interferometer.

$T_{\rm loss}$ = Noise contribution due to losses in the feed and input waveguide;

$T_{\rm cal}$ = Noise contribution due to injected calibration signal. The VLA noise source has a 50% duty cycle, so $T_{\rm cal}$ is one-half of the actual calibration value; and

$T_{\rm rx}$ = Receiver noise temperature measured at the room-temperature input flange to the receiver, including the contribution from the second and following stages.

Table 7–1 gives typical values at the zenith for these noise contributions for the six VLA receivers. For the VLA, the $T_{\rm spill}$ and $T_{\rm loss}$ contributions are somewhat higher than is usual for low-noise receivers because of the compromises that were made to have all receivers and feeds available simultaneously.

Table 7–1. Noise Contributions in the VLA Receivers

Wavelength Band	$T_{\rm bg}$ (K)	$T_{\rm sky}$ (K)	$T_{\rm spill}$ (K)	$T_{\rm loss}$ (K)	$T_{\rm cal}$ (K)	$T_{\rm rx}$ (K)	$T_{\rm sys}$ (K)
92 cm	25	3	15	7	5	70	125
20 cm	3	3	14	8	2	30	60
6 cm	3	3	7	5	2	30	50
3.6 cm	3	2	5	2	2	16	30
2 cm	3	8	6	13	6	80	116
1.3 cm	3	17	6	21	7	100	154

Several of these terms will vary with the position of the antenna. Obviously, the contribution from the galactic background depends upon the galactic coordinates being observed. $T_{\rm spill}$ will change as the orientation of the antenna with respect to the ground varies. More significantly, at high frequencies the contribution $T_{\rm sky}$ from atmospheric emission varies with time and antenna position. Assuming a plane-parallel atmosphere with temperature $T_{\rm atm}$, the dependence of $T_{\rm sky}$ on position is given by

$$T_{\rm sky} = T_{\rm atm}\left(1 - e^{-\tau \csc E}\right) , \tag{7–5}$$

where τ is the zenith attenuation and E is the elevation. Atmospheric attenuation also reduces the observed antenna temperature T_a from the T_{a_0} which would be measured outside the atmosphere by

$$T_a = T_{a_0} e^{-\tau \csc E} . \tag{7–6}$$

In this case the quantity of interest is the effective system temperature $T_{\rm eff}$ corrected for the effect of atmospheric attenuation,

$$T_{\rm eff} = T_{\rm sys} e^{\tau \csc E} . \tag{7–7}$$

At the VLA the effects of atmospheric attenuation are most serious at 1.3 cm wavelength where typical values of τ range between 0.03 and 0.17 (Spangler 1982).

3. Sensitivity of a Two-Antenna, Single-Multiplier, Correlation Interferometer

The two-antenna, single-multiplier, correlation interferometer is the basic element of a synthesis array, and in this section we consider the sensitivity of this basic element.

Several authors have analyzed the sensitivity of the simple two-antenna interferometer or the related correlation receiver, including Christiansen and Högbom (1985), Crane (1982), Rogers (1968, 1976), Staelin (1974), Thompson *et al.* (1986), Tiuri (1964), and Tiuri and Räisänen (1986). The following derivation follows that of Crane (1982).

Consider the case in which the interferometer shown in Figure 7–1 is observing an unpolarized point source of flux density S (S is measured in Janskys, $1\ \mathrm{Jy} = 10^{-26}\ \mathrm{W\,m^{-2}\,Hz^{-1}}$). The antenna temperature of antenna 1 due to the source is given by

$$T_{a_1} = \frac{\eta_{a1} A_1 S}{2k_B} = K_1 S\,, \tag{7–8}$$

where η_{a_1} is the aperture efficiency of antenna 1 (including the effect of losses) and A_1 is the geometrical area of antenna 1. The factor of 2 results from the fact that, since the source is unpolarized, a single-channel receiver on the antenna can accept only half of the power from the source. The expression for T_{a_2} is the same as Equation 7–8 with the subscript 1 replaced by 2. An important characteristic of the antenna is the sensitivity $K = T_a/S$ ($\mathrm{K\,Jy^{-1}}$) which, from Equation 7–8, is given by

$$K = \frac{\eta_a A}{2k_B}\,. \tag{7–9}$$

This quantity is a measure of the flux-collecting ability of the antenna. Table 7–2 shows typical values of K for the 25-m diameter shaped-reflector antennas of the VLA; for comparison the value of K for the 100-m telescope of the Max-Planck-Institut für Radioastronomie is $1.5\ \mathrm{K\,Jy^{-1}}$, and for the 1000-ft telescope of the Arecibo Observatory, between 6 and $15\ \mathrm{K\,Jy^{-1}}$, depending on frequency.

Table 7–2. Flux-Collecting Ability of VLA Antennas

Wavelength Band	η_a (%)	$K = T_a/S$ ($\mathrm{K\,Jy^{-1}}$)
92 cm	40	0.071
20 cm	51	0.091
6 cm	65	0.116
3.6 cm	62	0.110
2 cm	52	0.093
1.3 cm	43	0.082

The voltages as functions of time, $V(t)$, at the inputs to the multiplier are given by

$$\begin{aligned} V_1(t) &= S_1(t) + n_1(t)\,, \\ V_2(t) &= S_2(t) + n_2(t)\,, \end{aligned} \tag{7–10}$$

where $S(t)$ is the voltage due to the radio source and $n(t)$ is the voltage due to the system noise. The correlator multiplies $V_1(t)$ and $V_2(t)$ together and averages the product for some finite integration time. For this analysis we assume that the source is at the phase center of the interferometer, that the fringes have been stopped, and that time delays have been introduced so that $S_1(t)$ and $S_2(t)$ arrive at the multiplier in time synchronism (see Lecture 2). In this case, the correlator will produce a DC output resulting from the product $S_1(t)S_2(t)$, which corresponds to the desired measure of correlated power. An undesired, but unavoidable, zero-mean, time-varying output due to the various cross products in the multiplier will be superimposed on the DC output. To determine the sensitivity of the instrument, we wish to find the ratio of the DC output to the r.m.s. value of the time-varying component. Our approach will be to determine the power spectra of the various products generated in the correlator by using the Wiener–Khinchine theorem (Middleton 1960, p. 405). In this application the theorem states that the power spectrum of the product produced by the multiplier is equal to the Fourier transform of the autocorrelation function of the product. The various power spectra will then be integrated over the bandpass of the integrator to determine the power in the various terms. Several simplifying assumptions will be made. Assume that both receivers have identical frequency responses and, further, that $G_1 = G_2 = G$. Since both the signal and noise are Gaussian and white, the former assumption implies that the autocorrelation functions of $S_1(t)$ and $n_1(t)$ will have the same forms as the autocorrelations of $S_2(t)$ and $n_2(t)$, respectively.

The autocorrelation function, $\phi_{s_1}(\tau)$, of the signal at input 1 to the multiplier is given by

$$\phi_{s_1}(\tau) = \langle S_1 S_1' \rangle \,, \tag{7–11}$$

where $S_1 \equiv S_1(t)$, $S_1' \equiv S_1(t+\tau)$, and the angle brackets are used to denote an expected value. $S_1(t)$ and $S_2(t)$ differ only by a multiplicative constant

$$S_2(t) = \sqrt{\frac{K_1}{K_2}}\, S_1(t) \,. \tag{7–12}$$

Note that $\phi_{s_1}(0)$, the r.m.s. power contained in $S_1(t)$, is given by

$$\phi_{s_1}(0) = G k_B T_{a_1} \Delta\nu \,. \tag{7–13}$$

The autocorrelation function, $\phi_{n_1}(\tau)$, of the noise is given by

$$\phi_{n_1}(\tau) = \langle n_1 n_1' \rangle \,, \tag{7–14}$$

where $n_1 \equiv n_1(t)$ and $n_1' \equiv n_1(t+\tau)$. Note that the noise power at input 1 to the multiplier is given by

$$\phi_{n_1}(0) = G k_B T_{\mathrm{sys}_1} \Delta\nu \,, \tag{7–15}$$

and that the autocorrelation function of $n_2(t)$ differs only by a multiplicative constant from $\phi_{n_1}(\tau)$,

$$\langle n_2 n_2' \rangle = \frac{T_{\mathrm{sys}_2}}{T_{\mathrm{sys}_1}} \phi_{n_1}(\tau) \,. \tag{7–16}$$

Now, the autocorrelation function, $\phi_m(\tau)$, of the multiplier output is given by

$$\phi_m(\tau) = a^2 \langle V_1(t)V_2(t)V_1(t+\tau)V_2(t+\tau)\rangle . \tag{7–17}$$

Using Equation 7–10

$$\phi_m(\tau) = a^2 \langle (S_1 + n_1)(S_2 + n_2)(S_1' + n_1')(S_2' + n_2')\rangle , \tag{7–18}$$

where a is a constant.

Equation 7–18 can be expanded using a relationship from statistics which gives the expansion of the expectation of the product of four jointly Gaussian random variables (Davenport and Root 1958)

$$\begin{aligned}\phi_m(\tau) = a^2[&\langle (S_1 + n_1)(S_2 + n_2)\rangle \cdot \langle (S_1' + n_1')(S_2' + n_2')\rangle \\ &+\langle (S_1 + n_1)(S_1' + n_1')\rangle \cdot \langle (S_2 + n_2)(S_2' + n_2')\rangle \\ &+\langle (S_1 + n_1)(S_2' + n_2')\rangle \cdot \langle (S_2 + n_2)(S_1' + n_1')\rangle] .\end{aligned} \tag{7–19}$$

The required multiplications in Equation 7–19 can now be carried out simply. Setting the averages of the products of all uncorrelated voltages to zero, and using Equations 7–11, 7–12, 7–14 and 7–16, Equation 7–19 becomes

$$\begin{aligned}\phi_m(\tau) = a^2 \Bigg(&\frac{K_2}{K_1}\phi_{s_1}^2(0) + 2\frac{K_2}{K_1}\phi_{s_1}^2(\tau) \\ &+ \left(\frac{K_2}{K_1} + \frac{T_{\mathrm{sys}_2}}{T_{\mathrm{sys}_1}}\right)\phi_{s_1}(\tau)\phi_{n_1}(\tau) + \frac{T_{\mathrm{sys}_2}}{T_{\mathrm{sys}_1}}\phi_{n_1}^2(\tau)\Bigg) .\end{aligned} \tag{7–20}$$

Applying the Wiener–Khinchine theorem to Equation 7–20 and noting that the Fourier transform of a product is equal to the convolution of the Fourier transforms, we find that the power spectrum $\Phi_m(\nu)$ of the multiplier output is given by

$$\begin{aligned}\Phi_m(\nu) = a^2 \Bigg(&\frac{K_2}{K_1}\phi_{s_1}^2(0)\delta(\nu) + 2\frac{K_2}{K_1}\int_{-\infty}^{\infty}\Phi_{s_1}(\alpha)\Phi_{s_1}(\nu-\alpha)\,d\alpha \\ &+ \left(\frac{K_2}{K_1} + \frac{T_{\mathrm{sys}_2}}{T_{\mathrm{sys}_1}}\right)\int_{-\infty}^{\infty}\Phi_{s_1}(\alpha)\Phi_{n_1}(\nu-\alpha)\,d\alpha \\ &+\frac{T_{\mathrm{sys}_2}}{T_{\mathrm{sys}_1}}\int_{-\infty}^{\infty}\Phi_{n_1}(\alpha)\Phi_{n_1}(\nu-\alpha)\,d\alpha\Bigg) ,\end{aligned} \tag{7–21}$$

where $\delta(\nu)$ is the unit impulse at $\nu = 0$.

The power spectral components of the multiplier output are shown in Figure 7–2, taken from Crane (1982), for signal and noise with flat spectra passing through filters with rectangular passbands of width $\Delta\nu$. A physical interpretation of Figure 7–2 is useful. A component of $S_1(t)$ at a given frequency has multiplied the component of $S_2(t)$ at the same frequency to form $|S(t)|^2$; that is, $S(t)$ has been rectified and provides the desired DC output from the multiplier.

Figure 7–2. Power spectra for a two-antenna, single-multiplier, correlation interferometer.

Different frequency components of $S_1(t)$ and $S_2(t)$ beat together to form sum and difference frequencies. Since they are uncorrelated, frequency components of signal and noise, or of noise and noise, do not produce a DC signal when multiplied together but do produce the sum and difference frequencies.

The output of the multiplier is integrated by passing it through a low-pass filter. The power in the DC component is

$$P_{\text{noise}} = a^2 \frac{K_2}{K_1} \phi_s^2(0) \,. \tag{7–22}$$

Thus, from Equations 7–13 and 7–8 the DC voltage is

$$\langle V_0 \rangle = aGk_B S \Delta\nu \sqrt{K_1 K_2} \,. \tag{7–23}$$

To determine the fluctuations about this average value we must multiply $\Phi_m(\nu)$ by the response of the output filter, $|H(\nu)|^2$, and integrate. The power in the fluctuating component of the filter output is

$$\langle V_0^2(t) \rangle = \int_{-\infty}^{\infty} \Phi'_m(\nu) |H(\nu)|^2 \, d\nu \,, \tag{7–24}$$

where $\Phi'_m(\nu) = \Phi_m(\nu) - a^2 \frac{K_2}{K_1} \phi_s^2(0)\delta(\nu)$. We will consider an output filter which is an ideal integrator having integration time Δt. For practical values of Δt, $H(\nu)$ will be negligible except near $\nu = 0$, so $\Phi'_m(\nu)$ may be replaced by its value near $\nu = 0$, which is shown in Figure 7–2. For an ideal integrator the impulse response is

$$h(t) = \begin{cases} \frac{1}{\Delta t} \,, & 0 < t < \Delta t \,, \\ 0 \,, & \text{elsewhere} \,, \end{cases} \tag{7–25}$$

and the output noise power P_{noise} is, from Equation 7–24,

$$P_{\text{noise}} = \Phi'_m(0) \int_{-\infty}^{\infty} |H(\nu)|^2 \, d\nu \,. \tag{7–26}$$

Using Parseval's theorem

$$P_{\text{noise}} = \Phi'_m(0) \int_{-\infty}^{\infty} h^2(t) \, dt \tag{7–27}$$

$$= \frac{\Phi'_m(0)}{\Delta t} \tag{7–28}$$

$$= (aGk_B)^2 \frac{\Delta\nu}{\Delta t} \left(K_1 K_2 S^2 + \frac{K_1 S T_{\text{sys}_1}}{2} \left(\frac{T_{\text{sys}_2}}{T_{\text{sys}_1}} + \frac{K_2}{K_1} \right) + \frac{T_{\text{sys}_1} T_{\text{sys}_2}}{2} \right) . \tag{7–29}$$

The r.m.s. variation in the output is just $\sqrt{P_{\text{noise}}}$. In terms of flux density at the input, the r.m.s. noise ΔS at the filter output is

$$\Delta S = \frac{\text{r.m.s. noise (V)}}{\text{correlator scaling factor (V Jy}^{-1})} = \sqrt{P_{\text{noise}}} \bigg/ \frac{\partial V_0}{\partial S} \ . \tag{7–30}$$

Thus from Equations 7–30, 7–29, and 7–23,

$$\Delta S = \frac{1}{\sqrt{\Delta t \Delta\nu}} \sqrt{S^2 + \frac{S}{2}\left(\frac{T_{\mathrm{sys}_1}}{K_1} + \frac{T_{\mathrm{sys}_2}}{K_2}\right) + \frac{T_{\mathrm{sys}_1} T_{\mathrm{sys}_2}}{2K_1 K_2}}\,. \tag{7–31}$$

In the case of identical antennas and receivers, Equation 7–31 simplifies to

$$\Delta S = \frac{1}{\sqrt{\Delta t \Delta\nu}} \sqrt{S^2 + \frac{S T_{\mathrm{sys}}}{K} + \frac{T_{\mathrm{sys}}^2}{2K^2}}\,. \tag{7–32}$$

In the usual weak source case when $S \ll T_{\mathrm{sys}}/K$,

$$\Delta S = \frac{T_{\mathrm{sys}}}{K} \frac{1}{\sqrt{2\Delta t \Delta\nu}} = \frac{\sqrt{2} k_B T_{\mathrm{sys}}}{\eta_a A \sqrt{\Delta t \Delta\nu}}\,. \tag{7–33}$$

The corresponding expression for a total-power radiometer attached to an antenna identical to *one* of the interferometer antennas is $\sqrt{2}$ worse than this (i.e., ΔS is $\sqrt{2}$ bigger). Thus the sensitivity of a single correlation interferometer is $\sqrt{2}$ worse than a single antenna with the same *total* collecting area. The reason for this is that the interferometer does not make use of the information in the autocorrelation of the signals from each antenna separately.

In the case of a strong source, $S \gg T_{\mathrm{sys}}/K$, Equation 7–32 becomes

$$\Delta S = \frac{S}{\sqrt{\Delta t \Delta\nu}} = \frac{2 k_B T_a}{\sqrt{\Delta t \Delta\nu}\, \eta_a A}\,. \tag{7–34}$$

Equation 7–34 is the same expression as is obtained for a single antenna with a total-power radiometer observing a strong source. Notice that the sensitivity for a strong source is independent of antenna collecting area so that no improvement in sensitivity can be obtained by increasing the size or number of antennas. Equation 7–34 indicates that a noise source of a given noise temperature that is correlated between antennas will produce $\sqrt{2}$ larger noise fluctuations at the correlator output than will a noise source of the same noise temperature that is uncorrelated between antennas.

Expression 7–32 was derived for the special case of an observation of a point source with zero delay and phase, in which all the power received from the source is correlated between the two antennas. In the less ideal case the source is resolved with frequency-dependent structure and the properties of the system itself are also frequency-dependent. The quantity of interest is the correlated flux density

$$S_c(\nu) = A(\nu) \cos(\phi(\nu) + \phi_1(\nu) - \phi_2(\nu) + \omega\Delta\tau)\,, \tag{7–35}$$

where A, ϕ are the amplitude and phase of the visibility function, ϕ_i is the phase of the complex voltage gain of antenna i, and $\Delta\tau$ is the net delay. The DC voltage becomes

$$V_0 = a k_B \sqrt{K_1 K_2} \int_0^\infty G_1(\nu) G_2(\nu) S_c(\nu)\, d\nu\,, \tag{7–36}$$

and the output noise power is given by

$$P_{\text{noise}} = \frac{a^2 k_B^2 K_1 K_2}{2\Delta\tau} \int_0^\infty G_1^2(\nu) G_2^2(\nu) \left[S_c^2(\nu) + S^2(\nu) \right.$$
$$\left. + S(\nu) \left(\frac{T_{\text{sys}_1}(\nu)}{K_1} + \frac{T_{\text{sys}_2}(\nu)}{K_2} \right) + \frac{T_{\text{sys}_1}(\nu) T_{\text{sys}_2}(\nu)}{K_1 K_2} \right] d\nu \,. \quad (7\text{–}37)$$

In terms of correlated flux density at the input, the r.m.s. noise, ΔS, at the filter output is

$$\Delta S = \left\{ \int_0^\infty G_1^2(\nu) G_2^2(\nu) \left[S_c^2(\nu) + S^2(\nu) + S(\nu) \left(\frac{T_{\text{sys}_1}(\nu)}{K_1} + \frac{T_{\text{sys}_2}(\nu)}{K_2} \right) \right. \right.$$
$$\left. \left. + \frac{T_{\text{sys}_1}(\nu) T_{\text{sys}_2}(\nu)}{K_1 K_2} \right] d\nu \right\}^{\frac{1}{2}} \Big/ \sqrt{2\Delta t} \int_0^\infty G_1(\nu) G_2(\nu) \, d\nu \,. \quad (7\text{–}38)$$

This shows that the fluctuations in the correlated signal are independent of those in the total signal and must be added quadratically to the other noise terms. Also, in principle, the noise will differ between the real and imaginary parts of the visibility function and as the overall amplitude of the visibility function itself varies.

For a strong, fully resolved source ($A = 0$), once again making the usual simplifying assumptions, the r.m.s. noise is

$$\Delta S = \frac{S}{\sqrt{2\Delta\nu\Delta t}} \,. \quad (7\text{–}39)$$

This means that for those few experiments where an observer wishes to detect a weak spectral line or point source in the presence of a strong, extended continuum source whose flux density dominates the system temperature, the interferometer can offer a $\sqrt{2}$ improvement in sensitivity over a single antenna. Note that this improvement is in addition to any improvement in dynamic range which the interferometer is able to achieve by resolving the strong, extended source.

Expression 7–33 shows the dependence of the sensitivity of a correlation interferometer on the most important factors: system temperature, integration time, bandwidth and effective collecting area. Several other factors will affect the sensitivity by the order of a few percent to a few tens of percent. The most important of these effects we will call correlator efficiency,

$$\eta_c = \frac{\text{sensitivity of the correlator}}{\text{sensitivity of a perfect analog correlator having the same } \Delta t} \,. \quad (7\text{–}40)$$

η_c is needed because of the current tendency to use digital correlators. The correlator efficiency for a one-bit digital correlator of the type used in Mark II and Mark III VLBI is 64%, and for a three-level correlator, of the type used at the VLA, η_c is 81% (Cooper 1970).

A second effect, present in interferometers which have time-shared communication systems or digital correlators (especially those with recirculators), is a loss of integration time. For example, the VLA in normal continuum observing mode spends only 96.2% of observing time carrying out useful integration, and in the narrow-bandwidth spectral-line modes only 90.6% of observing time is useful (Escoffier 1979). Thus for the VLA in continuum mode η_c is 0.79 and in narrow-bandwidth spectral-line mode, 0.77.

Thus, the r.m.s. noise for a two-antenna, single-multiplier, correlation interferometer observing weak sources is given by

$$\Delta S = \frac{\sqrt{2} k_B T_{\rm sys}}{\sqrt{\Delta t \Delta \nu} A \eta_a \eta_c} . \tag{7–41}$$

Table 7–3 shows ΔS for the six observing bands of the VLA with η_c appropriate for continuum observing, Δt of 10 sec, $\Delta \nu$ of 46 MHz, using a single multiplier and one IF.

Table 7–3.
VLA Single-Multiplier Interferometer Sensitivity

Wavelength Band	ΔS (mJy)
92 cm	73
20 cm	28
6 cm	18
3.6 cm	11
2 cm	52
1.3 cm	84

4. Sensitivity of a Two-Antenna, Complex, Correlation Interferometer

The results of the preceding section were derived assuming the source of emission was unresolved and was located at the phase-tracking and delay-tracking centers of the interferometer. In general this will not be the case and a so-called "complex correlator" is used which has two multipliers, one of which has the signal from one antenna phase-shifted by 90°. The fringe patterns for the two correlators are phase-shifted by one-quarter of a fringe on the sky, and the flux density and phase of the visibility function for a source of arbitrary position and structure can be determined by combining the two measurements. The two outputs from the correlator are called the cosine and sine, or real and imaginary outputs. We will use the latter terminology and call the outputs of the correlator, calibrated in units of flux density, S_R and S_I for real and imaginary outputs, respectively. The measured amplitude, S_m, and the measured phase, ϕ_m, are determined by

$$S_m = \sqrt{S_R^2 + S_I^2} , \tag{7–42}$$

$$\phi_m = \tan^{-1} \frac{S_I}{S_R} . \tag{7–43}$$

Both S_R and S_I are contaminated by noise with r.m.s. value ΔS given by Equation 7–41, and we wish to determine noise estimates for S_m and ϕ_m. Notice that, in general, noise estimates for S_m and ϕ_m are not needed because synthesis images are computed directly from S_R and S_I. However, for completeness, we include the appropriate analysis here. The pioneering work was by S. O. Rice (1944, 1945, 1948), so that the basic distribution is often referred to as the *Rice distribution*. More recently, several authors have examined the problem in an astronomical context (Rogers 1968, 1976; Vinokur 1965; Hjellming and Basart 1982; Berge 1965; Moran 1973, 1976; Thompson *et al.* 1986).

The probability distribution of S_m, $P(S_m)$, is given by (Hjellming and Basart 1982)

$$P(S_m) = \frac{S_m}{\Delta S^2}\, I_0\left(\frac{S_m S}{\Delta S^2}\right) \exp\frac{-(S_m^2 + S^2)}{2\Delta S^2}\,, \tag{7–44}$$

where I_0 is the modified Bessel function of the first kind, order zero, and S is the true amplitude. Plots of $P(S_m)$, adapted from Hjellming and Basart (1982), are shown in Figure 7–3 for various values of $S/\Delta S$. For small values of $S/\Delta S$, $P(S_m)$ is close to a Rayleigh distribution and S_m is a biased estimator of S (Moran 1976):

$$\langle S_m \rangle \approx \sqrt{\frac{\pi}{2}}\,\Delta S\left(1 + \frac{S^2}{4\Delta S^2}\right)\,, \tag{7–45}$$

and

$$\sigma_{S_m} \approx \sqrt{2 - \frac{\pi}{2}}\,\Delta S\left(1 + \frac{S^2}{4\Delta S^2}\right)\,. \tag{7–46}$$

For large values of $S/\Delta S$,

$$I_0\left(\frac{S_m S}{\Delta S^2}\right) \approx \frac{\Delta S}{\sqrt{2\pi S_m S}}\,\exp\frac{S_m S}{\Delta S^2}\,, \tag{7–47}$$

and

$$P(S_m) \approx \frac{1}{\Delta S}\sqrt{\frac{S_m}{2\pi S}}\,\exp\frac{-(S_m - S)^2}{2\Delta S^2}\,. \tag{7–48}$$

Since we need only consider $S_m \approx S$,

$$P(S_m) \approx \frac{1}{\sqrt{2\pi}\Delta S}\,\exp\frac{-(S_m - S)^2}{2\Delta S^2}\,, \tag{7–49}$$

which is a Gaussian distribution with standard deviation ΔS.

The probability distribution of the phase error $\phi - \phi_m$, where ϕ is the true phase, is given by (Hjellming and Basart 1982)

$$P(\phi - \phi_m) = \frac{1}{2\pi}\,\exp\left(\frac{-S^2}{2\Delta S^2}\right)\left(1 + G\sqrt{\pi}e^{G^2}(1 + \operatorname{erf} G)\right)\,, \tag{7–50}$$

Figure 7–3. The probability distribution of the measured amplitude is plotted as a function of the apparent signal-to-noise ratio for a number of values of the true signal-to-noise ratio.

where erf is the error function and $G(\theta) = \frac{S\cos\theta}{\sqrt{2}\Delta S}$. Plots of $P(\phi - \phi_m)$, reproduced from Hjellming and Basart (1982), are shown in Figure 7–4 for several values of $S/\Delta S$. For small $S/\Delta S$ the r.m.s. phase error is given by (Moran 1976)

$$\sigma_{\phi_m} \approx \frac{\pi}{\sqrt{3}} \left(1 - \sqrt{\frac{9}{2\pi^3}} \, \frac{S}{\Delta S} \right) , \qquad (7\text{–}51)$$

while for large values of $S/\Delta S$, $P(\phi - \phi_m)$ approaches a Gaussian distribution with standard deviation

$$\sigma_{\phi_m} = \frac{\Delta S}{S} \, . \qquad (7\text{–}52)$$

Figure 7–4 demonstrates clearly why an observer, who is trying to establish if correlated signal is present on a given interferometer pair by looking at a time sequence of the amplitude and phase, should look at the phase rather than the amplitude. The difference between the phase distributions for values of $S/\Delta S$ equal to 0 and 1 is much more obvious than the difference between the associated amplitude distributions.

5. Sensitivity of a Synthesis Array to a Point Source

The brightness distribution $I(l,m)$ of a source is determined from the complex visibility function $V(u,v)$ using (from Lecture 1, Eq. 1–9)

$$I(l,m) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} V(u,v) e^{2\pi i (ul+vm)} \, du \, dv \, , \qquad (7\text{–}53)$$

Figure 7–4. The probability distribution of the measured phase is plotted as a function of $\phi - \phi_m$ for a number of values of the true signal-to-noise ratio.

where l and m are direction cosines and u and v are baseline coordinates. V is related to the real and imaginary outputs of a complex correlator, described in the previous section, by

$$V = S_R + iS_I \,, \tag{7–54}$$

where $i^2 = -1$. We wish to determine the r.m.s. noise in $I(l, m)$ given that both S_R and S_I contain r.m.s. noise ΔS. In practice, and now including the tapering and weighting functions discussed in Lecture 6 (Section 2), Expression 7–53 is approximated by

$$I_m(l,m) = K \sum_{\ell=0}^{2L} T_\ell W_\ell V_\ell e^{2\pi i(u_\ell l + v_\ell m)} , \tag{7–55}$$

where I_m is the measured brightness distribution, K is a constant, L is the number of measurements of V plus a zero-spacing flux density ($\ell = 0$), and the factor 2 in the limit $2L$ is included because V is Hermitian so that, if $V(u,v)$ is measured, $V(-u,-v)$ is also known. u_ℓ and v_ℓ are the u and v coordinates of the ℓ^{th} measurement of V. By the appropriate choice of K, the units of $I_m(l,m)$ can be expressed as flux density per synthesized beam. Equation 7–55 can be rewritten using Equation 7–54

$$I_m(l,m) = 2K \sum_{\ell=1}^{L} T_\ell W_\ell \big(S_{R_\ell} \cos 2\pi(u_\ell l + v_\ell m) - S_{I_\ell} \sin 2\pi(u_\ell l + v_\ell m)\big) , \tag{7–56}$$

where it has been assumed, to simplify the expression, that no zero-spacing flux density is available. The easiest way to determine the noise in $I_m(l,m)$ is to consider the noise at the center of the image where the expressions become very simple. Equation 7–56 is a so-called 'direct Fourier transform', and the noise will be the same at all points on the image. At the image center

$$I_m(0,0) = 2K \sum_{\ell=1}^{L} T_\ell W_\ell S_{R_\ell} . \tag{7–57}$$

Now, for a point source of flux density S located at $l = m = 0$,

$$S_{R_\ell} = S + n_{R_\ell} , \tag{7–58}$$

where n_{R_ℓ} is the noise in the real part of the correlator output and has the properties $\langle n_{R_\ell} \rangle = 0$ and $\langle n_{R_\ell}^2 \rangle = \Delta S^2$, where ΔS is given by Equation 7–41. The expected value at the image center is

$$I_m(0,0) = 2KS \sum_{\ell=1}^{L} T_\ell W_\ell . \tag{7–59}$$

To express $I_m(l,m)$ as flux density per beam area, K is set equal to

$$1 \bigg/ 2 \sum_{\ell=1}^{L} T_\ell W_\ell \ ,$$

so that

$$I_m(0,0) = S , \tag{7–60}$$

and the r.m.s. noise in the image, ΔI_m, is

$$\Delta I_m = 2K\Delta S\sqrt{\sum_{\ell=1}^{L} T_\ell^2 W_\ell^2}\,. \tag{7–61}$$

For a naturally weighted, untapered image, this simplifies to

$$\Delta I_m = \frac{\Delta S}{\sqrt{L}}\,. \tag{7–62}$$

For an array with N antennas, N_b baselines ($N_b = N(N-1)/2$ if all antenna pairs are correlated), $N_{\rm IF}$ IFs, correlator integration time Δt, and total observing time T, the number of measurements is

$$L = N_b N_{\rm IF} T/\Delta t\,. \tag{7–63}$$

Combining Equations 7–41, 7–62, and 7–63 gives the desired expression for the noise in the image

$$\Delta I_m = \frac{\sqrt{2}k_B T_{sys}}{\eta_a \eta_c A\sqrt{N_b N_{\rm IF} T\Delta\nu}}\,. \tag{7–64}$$

Notice that at the image center only the noise from the real correlators affects the image. Elsewhere in the image the noise from both the real and imaginary parts will contaminate the image, but the noise estimate is still the same as given by Equation 7–64 because the real noise is weighted by $\cos 2\pi(ul+vm)$, the imaginary noise is weighted by $\sin 2\pi(ul+vm)$, and the real and imaginary noise terms are uncorrelated.

For a synthesis array of N antennas using two IF's, if all possible baselines are correlated, $N_b N_{\rm IF}$ approaches N^2 as N becomes large,

$$\Delta I_m = \frac{\sqrt{2}k_B T_{\rm sys}}{\eta_a \eta_c N A\sqrt{T\Delta\nu}}\,, \tag{7–65}$$

which, with η_c equal to 1, is the same noise as would be expected from a single antenna of collecting area NA with two IF's connected to total-power radiometers. That the large synthesis array has the same sensitivity as a single antenna of the same total area is not surprising because, as N becomes large, the fraction of the information lost by the synthesis array because it does not carry out the autocorrelations becomes negligible. The synthesis array has the advantage, however, that all points in the field of view are observed with sensitivity ΔI_m, while the single antenna must observe each point separately for time T.

Most modern synthesis arrays have the capability of operating as a phased-array in which the IF signals from all antennas are added together after the delay lines, to create an IF received by the synthesized beam at the array phase center. Such a phased-array output is useful for VLBI and for spectroscopic observations. In principle this output has all the information present in a single antenna of the same collecting area and resolution, so that even the autocorrelation information can be recovered. In practice, several effects may reduce the sensitivity of this output by a few tens of percent below the expected sensitivity. If the output is formed by adding together digitized outputs from a small number of antennas (less than 15, say), the sensitivity will be slightly less than expected and will vary depending on whether odd or even numbers of antennas are added together (Van Ardenne 1979, 1980). If the output is again digitized to allow VLBI recording or digital spectral analysis, the loss of sensitivity η_c occurs again. An effect present in the VLA phased-array output is that the phased-array IF is reconstructed using pulses of finite width rather than delta functions. This reduces the effective bandwidth of the output and lowers its sensitivity by a few percent.

Table 7–4 shows the theoretical ΔI_m for the six observing bands of the VLA for a twelve-hour synthesis with η_c appropriate for continuum observing, $\Delta\nu$ of 46 MHz, using two IF's. In practice, atmospheric attenuation, variations in aperture efficiency, the presence of radio-frequency interference, and many other factors will prevent one from reaching the theoretical ΔI_m.

Table 7–4.
Theoretical VLA Sensitivity
for a Twelve-Hour Synthesis

Wavelength Band	ΔI_m (μJy/beam)
92 cm	42
20 cm	16
6 cm	10
3.6 cm	7
2 cm	30
1.3 cm	48

In natural weighting, every correlator measurement is given the same weight. This gives the highest sensitivity for detecting point sources. As described in Lecture 6, tapering and weighting functions can be applied to each measurement to control, to some extent, the beam shape. Applying such functions degrades the point-source sensitivity (makes ΔI_m larger) by

$$\sqrt{\sum_{\ell=1}^{L} T_\ell^2 W_\ell^2} \Bigg/ \sum_{\ell=1}^{L} T_\ell W_\ell \,.$$

For an East–West synthesis array such as the Westerbork Synthesis Radio Telescope, the weighting function for uniform weighting is known analytically ($W \propto q$) and the point-source sensitivity can be determined analytically. Because the sampling function for the Very Large Array varies greatly with declination, integration time, observing strategy, and total observing time, for example, the weighting function for uniform weighting is not known analytically. Instead, as described in Lecture 6, the weight for each measurement is determined from the inverse of the local density of measurements, which usually is measured over one cell in the u-v plane, but the user can select a larger area.

Consequently, when only a few measurements are spread over many cells in the u-v plane, the local density of measurements for most u-v cells will be either zero or one and the sensitivity will be close to that for natural weighting. At the other extreme of many measurements spread over a few cells, the density of measurements near the center of the u-v plane will be very high (several hundred per cell), the weights low, and the sensitivity will be considerably degraded (by a factor of order 1.2 to 3).

Application of a Gaussian taper tends to cancel the effect of the weighting function for uniform weighting on the point-source sensitivity. The sensitivity will be best for an optimum value of the taper dispersion σ at which the tapering function best matches natural weighting, and will degrade monotonically in either direction for other values of σ.

The examples in Figure 7–5 illustrate the dependence of the point-source sensitivity upon the number of cells in the u-v plane and upon σ. The calculations were done for a twelve-hour synthesis at a declination of 90°, with an integration time of 100 seconds. In addition to the curve labeled "Natural" which shows the effect of combining natural weighting and a Gaussian tapering function, the other four curves show the effects of combining a Gaussian tapering function and uniform weighting, with the u-v plane spanned by the number of cells indicated. (The number in parentheses indicates the maximum number of measurements per u-v cell in each example.)

6. Sensitivity of a Synthesis Array to an Extended Source

A very important aspect of the sensitivity properties of a synthesis array is the difference between the sensitivity to point sources and to extended sources. The units of the brightness image can be expressed as Janskys per synthesized beam area, and the r.m.s. noise in the image is ΔI_m Janskys per synthesized beam. Suppose that the size of the synthesized beam is varied by scaling the size of the array. A point source of flux density S has a constant apparent brightness of S Janskys per synthesized beam, independent of the size of the synthesized beam. Therefore the signal-to-noise ratio of a point source, $S/\Delta I_m$, is independent of the beam size. For an extended source that is larger than the synthesized beam, having constant brightness I Janskys per steradian, the flux density per synthesized beam is $I\Omega_s$, where Ω_s is the area of the synthesized beam in steradians. Therefore, the signal-to-noise ratio for this source is

$$\frac{I\Omega_s}{\Delta I_m},$$

which improves as the synthesized beam is made larger, so long as the beam is smaller than the smallest detectable source structure and I_m does not increase significantly. Fomalont and Wright (1974) give further discussion of this point. In general, the resolution of an array increases linearly with the size of the array,

Figure 7–5. The effects of applying a Gaussian tapering function on the point-source sensitivity of the VLA, shown for natural weighting and for uniform weighting with four different numbers of cells spanning the u-v plane. The calculations are for a source at the North Celestial Pole, observed for twelve hours with an integration time of 100 seconds, and have been normalized to the result for untapered natural weighting.

but the sensitivity to extended structure decreases as the square of the size. Thus, the VLA observers who propose to observe an extended source in the **A** configuration after observing it in the **B** configuration and needing 3 times more resolution, must remember that they will have an order of magnitude less sensitivity to the extended structure (see also Lecture 24).

The sensitivity to extended structure, as well as being improved by scaling the array to match the size of the source, can usually be improved by applying a tapering function as described in Lecture 6. The primary improvement in sensitivity arises because the area of the synthesized beam Ω_s increases approximately as σ^{-2} but also because, as indicated by Figure 7–5, ΔI_m remains constant or even decreases for a wide range of σ. So for $\sigma \approx 0.2$–$0.3 q_{\max}$, the improvement in sensitivity to an extended source may be a factor of 4–10, compared with uniform weighting.

For a more detailed, although still qualitative, understanding of the sensitivity of the VLA to extended structure, consider the response to a circular

Figure 7–6. The effects of applying a Gaussian tapering function on the peak response of the VLA to a circular Gaussian source with $\theta = 2.08\lambda/q_{\max}$, shown for natural weighting and for uniform weighting with four different numbers of cells spanning the u-v plane. The calculations are for a source at the North Celestial Pole, observed for twelve hours with an integration time of 100 seconds, and have been normalized to the response to a point source.

Gaussian source with full width at half maximum θ. The average surface brightness, I_s, for such a source with a total flux density S is given by

$$I_s(S,\theta) = \frac{S}{1.133\theta^2}\,. \tag{7–66}$$

As with the point-source sensitivity, the peak response of the VLA to a circular Gaussian source depends upon many variables; Figure 7–6 shows the peak responses to a circular Gaussian source with a θ of $2.08\lambda/q_{\max}$ for the same examples used in Figure 7–5. The surface-brightness sensitivity $\Delta I_s(\theta,\sigma)$ for a source with angular size θ and imaged with a Gaussian tapering function of dispersion σ is then given by

$$\Delta I_s(\theta,\sigma) = \frac{\Delta I_m(\sigma) I_s(S,\theta)}{I_m(S,\theta,\sigma)}\,. \tag{7–67}$$

As illustrated in Figures 7–5 and 7–6, this function will depend upon the details of the observing strategy used, the parameters chosen for the image, and many other variables. However, for detecting weak, extended (i.e., low-surface-brightness) sources, natural weighting provides the greatest sensitivity. Figures 7–5 and 7–6 show that natural weighting has both the lowest ΔI_m and the greatest I_m which through Equation 7–67 imply the lowest ΔI_s. The dependence of the surface-brightness sensitivity for natural weighting on source size and taper dispersion is shown in Figure 7–7. The optimum taper dispersion which minimizes ΔI_s can be determined as a function of source size; that relation is shown by the diagonal line in Figure 7–7. The relation between the minimum ΔI_s and source size determines the surface-brightness sensitivity. The results for the four configurations of the VLA are shown in Figure 7–8. The more compact a configuration, the better is its sensitivity to extended emission. At the other extreme, as expected, all configurations have asymptotically the same sensitivity to small and unresolved sources. The upward curvatures of the four curves show the effects of missing measurements at the shortest spacings.

7. The Effects of Convolution and Gridding on Sensitivity

The sensitivity analyses in Section 5 are appropriate when images are made using a so-called 'direct Fourier transform' without convolution in the *u-v* plane, in which case the noise is uniform over the image. In most practical cases, images are made using the Fast Fourier transform which requires that the *u-v* plane data be convolved and gridded before being transformed. Analysis of the effect of these operations on the signal-to-noise ratio in the image is complicated and is discussed extensively in three reports (Greisen 1976, 1979; Clark 1976). The reader is referred to these reports for details; here we will attempt only to give the reader a physical understanding for the effect, which can significantly degrade the signal-to-noise at the edge of the image. The effect is caused by a combination of two processes; the aliasing of noise back into the image and the division of the image by the Fourier transform of the *u-v* plane convolving function to remove the effects of this convolution.

Consider only the noise in the *u-v* plane. If we compute the distribution of the noise in the image plane using a 'direct Fourier transform', the noise will have the same variance everywhere, as shown in Figure 7–9a (provided the smoothing effect of the correlator integration time is negligible). Now let us convolve the *u-v* plane data with a convolving function and again compute the image using a 'direct Fourier transform'. Now the distribution of the variance of the noise is shown in Figure 7–9b, where the variance in the image is multiplied by the Fourier transform of the autocorrelation function of the convolving function (which is equal to the square of the Fourier transform of the convolving function). Now, if the convolved *u-v* plane data are again sampled at points with normalized spacing $\frac{1}{2}$, and transformed using a Fast Fourier transform, the variance of the noise in the image is significantly changed by aliasing, as shown in Figure 7–9c. Finally, after the image has been divided by the Fourier transform of the *u-v* plane convolving function, the variance of the noise is as shown in Figure 7–9d. Clearly this final division process has enhanced the noise at the edge of the image, resulting in a degraded signal-to-noise ratio.

Greisen (1979) computes the size of this effect for many different convolving

Figure 7–7. The effects of tapering and source size on the surface-brightness sensitivity of the VLA, shown for natural weighting and assuming a Gaussian tapering function of dispersion σ. The calculations are for a source at the North Celestial Pole, observed for twelve hours with an integration time of 10 seconds. The contour levels are 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, and 5000 μJy per square arcsecond, calculated for an observation at 1465 MHz in the **A** configuration, $\Delta\nu$ of 46 MHz, using two IFs. The scaling to other configurations and frequencies is as $\nu^2 3.285^{2(1-n)}$ where $n = 1, 2, 3$, or 4 for the **A, B, C**, or **D** configuration, respectively. The diagonal line relates the source size and the corresponding optimum taper dispersion, and is given approximately by $\log_{10}(\sigma/q_{\rm max}) \approx -.72 - 1.01 \log_{10}(\theta/(\lambda/q_{\rm max}))$.

functions. The commonly used pillbox function with width equal to the grid spacing, for example, significantly degrades the sensitivity over most of the image, with the worst degradation being a factor of 0.4 decrease in signal-to-noise at the image corners. Other types of convolving functions can be found which only affect the outer one-quarter of the image, with a worst case degradation of a factor of 0.5.

8. Effect of Primary Beam on Sensitivity

Ignoring for the moment all image distortions except additive noise and the effect of the primary beam of the individual antennas comprising the synthesis

Figure 7–8. The surface-brightness sensitivities of the four configurations of the VLA from the calculations shown in Figure 7–7. The units of the ordinate are μJy per square arcsecond at 1465 MHz and scale to other frequencies as $\nu^2 \Delta S$.

Figure 7–9. The effects of convolution and gridding in the u-v plane on the noise in an image: **(a)** 'Direct Fourier transform', no convolution or gridding. **(b)** 'Direct Fourier transform' after convolution. **(c)** Fast Fourier transform after convolution and gridding. **(d)** Fast Fourier transform after convolution and gridding, followed by division by the Fourier transform of the convolving function.

array, we may express the measured brightness image $I_m(l,m)$ as

$$I_m(l,m) = I(l,m)P(l,m) + N(l,m)\,, \tag{7–68}$$

where $I(l,m)$ is the true brightness distribution, $P(l,m)$ is the response of the antenna primary beam and $N(l,m)$ is additive noise with r.m.s. value ΔI_m. If the variation of $P(l,m)$ across $I(l,m)$ is not negligible its effect may be removed by dividing the image by $P(l,m)$, in which case

$$I_m(l,m)/P(l,m) = I(l,m) + N(l,m)/P(l,m). \tag{7–69}$$

In this case Equation 7–69 shows that the noise is enhanced at the edge of the image, reducing the sensitivity in this region.

Acknowledgments

The authors thank F. Schwab and A. Bridle for their comments on the lecture, and thank R. Hjellming for providing Figures 7–3 and 7–4.

References

Berge, G. L. (1965), *An Interferometric Study of Jupiter's Decimeter Radio Emission*, Ph. D. Thesis, Caltech.

Bevington, P. R. (1969), *Data Reduction and Error Analysis for the Physical Sciences*, McGraw–Hill, New York.

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

Clark, B. G. (1976), "Gridding and signal-to-noise ratios", VLA Scientific Memorandum No. 124, NRAO.

Cooper, B. F. C. (1970), "Correlators with two-bit quantization", *Aust. J. Phys.*, **23**, 521–527.

Crane, P. C. (1982), "Signal analysis of a correlation interferometer", VLA Scientific Memorandum No. 140, NRAO.

Davenport, W. B. and Root, R. L. (1958), *An Introduction to the Theory of Random Signals and Noise*, McGraw–Hill, New York.

Escoffier, R. P. (1979), "Correlator system observers manual", VLA Technical Report No. 39, NRAO.

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

Greisen, E. W. (1976), "On the effects of convolution in the *u*-*v* plane", VLA Scientific Memorandum No. 123, NRAO.

Greisen, E. W. (1979), "The effects of various convolving functions on aliasing and relative signal-to-noise ratios", VLA Scientific Memorandum No. 131, NRAO.

Hjellming, R. M. and Basart, J. P. (1982), "The theory of the instrument", Chapter 2 in *An Introduction to the NRAO Very Large Array*, R. M. Hjellming, Ed., NRAO.

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

Middleton, D. (1960), *An Introduction to Statistical Communication Theory*, McGraw–Hill, New York.

Moran, J. M. (1973), "Spectral line analysis of very-long-baseline interferometric data", *Proc. IEEE*, **61**, 1236–1242.

Moran, J. M. (1976), "Very long baseline interferometric observations and data reduction", in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, pp. 228–260.

Rice, S. O. (1944), "Mathematical analysis of random noise", *Bell Sys. Tech. J.*, **23**, 282; (reprinted in *Selected Papers on Noise and Stochastic Processes*, N. Wax, Ed., Dover, New York, 1954).

Rice, S. O. (1945), "Statistical properties of random noise currents", *Bell Sys. Tech. J.*, **24**, 46; (reprinted in *Selected Papers on Noise and Stochastic Processes*, N. Wax, Ed., Dover, New York, 1954).

Rice, S. O. (1948), "Statistical properties of a sine wave plus random noise", *Bell Sys. Tech. J.*, **27**, 109.

Rogers, A. E. E. (1968), "Spectral line interferometry and interferometer noise analysis", Technical Report No. 441, MIT Lincoln Labs.

Rogers, A. E. E. (1976), "Theory of two-element interferometers", in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, pp. 139–157.

Spangler, S. R. (1982), "Correction of VLA K-band amplitudes for atmospheric attenuation", VLA Scientific Memorandum No. 143, NRAO.

Staelin, D. H. (1974), "The detection and measurement of radio astronomical signals", MIT Dept. Elec. Eng., Lecture Notes.

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

Tiuri, M. E. (1964), "Radio astronomy receivers", *IEEE Trans. Ant. Propagat.*, **AP-12**, 930–938.

Tiuri, M. E. and Räisänen, A. V. (1986), "Radio-telescope receivers", Chapter 7 in J. D. Kraus, *Radio Astronomy*, Second Edition, Cygnus-Quasar Books (Powell, Ohio).

Van Ardenne, A. (1979), "Loss of sensitivity with Westerbork as a digital tied array, a first study", Netherlands Foundation for Radio Astronomy, Note No. 289.

Van Ardenne, A. (1980), "Calibrating the tied array in the case of analog summation of analog signals", Netherlands Foundation for Radio Astronomy, Note No. 315.

Vinokur, M. (1965), "Optimization in the search for a sinusoid of known period in the presence of noise. Application to radio astronomy", *Ann. D'Ap.*, **28**, 412–445.

8. DECONVOLUTION

Tim Cornwell and Robert Braun

1. Deconvolution

This lecture describes how the visibility samples collected by an interferometric array can be used to produce a high quality image of the sky. As noted in Lecture 1, the image formed by simple Fourier transformation of the observed, sampled visibilities by the methods described in Lecture 6 will have defects due to the limited sampling of the u-v plane. Nonlinear deconvolution is required to correct these defects.

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:

$$V(u,v) = \iint_S I(l,m) e^{-2\pi i(ul+vm)}\, dl\, dm\,, \tag{8–1}$$

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 $\Delta l \le \frac{1}{2u_{\max}}$, $\Delta m \le \frac{1}{2v_{\max}}$, $N_l \Delta l \ge \frac{1}{u_{\min}}$, and $N_m \Delta m \ge \frac{1}{v_{\min}}$. 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

$$V(u_r, v_r) = \widehat{V}(u_r, v_r) + \epsilon(u_r, v_r)\,, \tag{8–3}$$

where $\epsilon(u_r, v_r)$ is a complex, normally distributed random error due to receiver noise, and r 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

$$V(u, v) = W(u, v)\big(\widehat{V}(u, v) + \epsilon(u, v)\big)\,, \tag{8–4}$$

where $W(u, v)$ is a weighted sampling function (see Lecture 6, Eq. 6–8) which is non-zero only for sampled points of the u-v plane,

$$W(u, v) = \sum_r W_r \delta(u - u_r, v - v_r)\,. \tag{8–5}$$

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

$$I^D_{p,q} = \sum_r W(u_r, v_r)\,\mathrm{Re}\left(V(u_r, v_r) e^{2\pi i(pu_r\Delta l + qv_r\Delta m)}\right) \tag{8–7}$$

and

$$B_{p,q} = \sum_r W(u_r, v_r)\,\mathrm{Re}\left(e^{2\pi i(pu_r\Delta l + qv_r\Delta m)}\right)\,. \tag{8–8}$$

$E_{p,q}$ in Equation 8–6 is the noise image obtained by replacing V in Equation 8–7 by $\epsilon(u_r, v_r)$. 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 $W(u_r, v_r)$ to the reciprocal of the variance of the error in $V(u_r, v_r)$ will optimize the signal-to-noise ratio in the final image, whereas setting it to the reciprocal of some approximation of the local density of samples will minimize the sidelobe level (see Lecture 6).

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 $I' = A * D$ 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 and Hunt 1977) will not work: a nonlinear procedure is required.

Interferometrists call the homogeneous solutions "invisible distributions" (Bracewell and 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. Some of the more common shortcomings in visibility sampling and their signatures in the principal solution are illustrated in Figure 8–1. 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 now 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 a radio source by 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', usually Gaussian, beam 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) Go to (1) unless any remaining peak is below some user-specified level.

(4) 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).

(5) Add the residuals of the dirty image to the 'CLEAN' image.

The fifth 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

Figure 8–1. Problems with the dirty image. Some of the more common shortcomings in visibility sampling and their signatures in the dirty image are illustrated in panels **(a)**-**(d)**. Panel **(e)** gives an example of a visibility sampling and a corresponding invisible distribution.

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 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 and Schwab (Schwab 1984b, top right corner of p. 1078) 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 'CLEAN'ed 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 and 14).

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 'CLEAN'ing 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 program 'MX'.

2.4. Other related algorithms.

Several algorithms have been invented with the aim of correcting some deficiencies of 'CLEAN'.

Steer, Dewdney and 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 seems to be considerable compared to 'CLEAN', but the radio astronomy community has little experience with this variant of the algorithm so its ability to handle different practical situations is not yet well-known.

Segalovitz and 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 to ameliorate partially the striping instability to which 'CLEAN' is susceptible (see Sec. 3.7 below).

3. Practical Details and Problems of 'CLEAN' Usage

Theoretical understanding of 'CLEAN' is relatively poor even though the original algorithm is about 15 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. Clearly, 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. 'CLEAN'ing close to the edge of a dirty image computed by an FFT is particularly risky.

Most of our understanding of 'CLEAN' 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.

3.1. 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 'CLEAN'ing 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 6) 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 'CLEAN'ed.

3.2. Number of iterations and the loop gain.

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}). 'CLEAN'ing deep into the noise is usually a waste of time unless you specifically wish to analyze the extended, low surface-brightness emission (but see Sec. 3.4 below).

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.

3.3. 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 'CLEAN'ing, 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 'CLEAN'ing of the dirty image; one can experiment by simply increasing N_{CL}. Ultimately, it may actually be necessary to measure the appropriate data!

3.4. The 'CLEAN' beam.

The 'CLEAN' 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 most common method of choosing the 'CLEAN' beam is to fit an elliptical Gaussian to the central region of the dirty beam. One should remember that this choice is merely the result of a compromise between resolution and apparent image quality and that larger or smaller beams may be appropriate in particular cases. If one is prepared to tolerate a decrease in the apparent quality of the 'CLEAN' image, and if both the signal-to-noise ratio and the u-v coverage are good, then often a smaller 'CLEAN' beam can be used.

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. Provided that 'CLEAN' is run to convergence, the integral of the 'CLEAN' image will often provide an accurate estimate of the flux density of an extended object, 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.

3.5. 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 'CLEAN'ing 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.

3.6. 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 'CLEAN'ing strategies upon the final image.

3.7. 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 'CLEAN'ing. 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. If 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.

4. 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. There is no consensus on this subject evident yet in the literature (e.g., Frieden 1972; Wernecke and D'Addario 1976; Gull and Daniell 1978; Jaynes 1982; Narayan and Nityananda 1984, 1986; Cornwell and 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" (see Cornwell 1984a). 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 and 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:

$$\chi^2 = \sum_r \frac{\left|V(u_r, v_r) - \widehat{V}(u_r, v_r)\right|^2}{\sigma^2_{V(u_r, v_r)}}. \tag{8–10}$$

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 and Evans 1985).

Algorithms for solving this maximization problem have been given by Wernecke and D'Addario (1976), by Cornwell and Evans (1985), and by Skilling and Bryan (1984). The Cornwell–Evans algorithm is coded in NRAO's Astronomical Image Processing System (AIPS) as 'VM'. It is generally faster than 'CLEAN' for larger images; the break-even point being for images of about 1 million pixels.

5. Practical Details of the Use of MEM

The following description relates to the AIPS MEM algorithm, 'VM'.

5.1. 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. 5.5 below).

5.2. 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.

5.3. 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 and 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.

5.4. 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.

5.5. 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.

6. Comparison of 'CLEAN' and MEM

'CLEAN' has dominated deconvolution in radio astronomy since its invention nearly 15 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 slightly 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'.

7. Other Methods, Including Hybrids

Deconvolution in radio astronomy is currently 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.

The concept of a default image can be extended to 'CLEAN' and other algorithms, and it will probably improve their performance and suggest different types of algorithm.

A relatively unexplored area is that of *linear* methods with boundary conditions, such as singular value decomposition (SVD; e.g., Andrews and Hunt 1977). SVD is a generalization of eigenfunction analysis to systems split into two domains, such as the sky and the u-v planes. Using SVD, the constraint of confinement could be applied to estimate unsampled data and thus remove sidelobes. Unfortunately, it is very expensive to use unless the geometry of the imaging system is simple in some way and thus it may be applicable only to certain telescopes, such as East–West arrays.

A method proposed by Braun and Walterbos (1985) addresses the problem of incomplete short spacing information in the absence of other shortcomings in the visibility sampling. A least-squares fit to a matched functional form is used to provide an analytic continuation of the background beneath the locations of extended sources. The technique is efficient and successful for this restricted problem in cases where the confinement constraint can be effectively applied.

Increasing use is being made of hybrid techniques which 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 'CLEAN'ing 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 'CLEAN'ing 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. Such techniques offer considerable promise to general application, especially if their use can be streamlined.

It is ironic that, formally, more is known about the type of images generated by MEM than by 'CLEAN' (see e.g., Narayan and 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. and Hunt, B. R. (1977), *Digital Image Restoration*, Prentice–Hall (Englewood Cliffs, NJ).

Bracewell, R. N. and Roberts, J. A. (1954), "Aerial smoothing in radio astronomy", *Aust. J. Phys.*, **7**, 615–640.

Braun, R. and Walterbos, R. A. M. (1985), "A solution to the short spacing problem in radio interferometry", *Astron. Astrophys.*, **143**, 307–312.

Clark, B. G. (1980), "An efficient implementation of the algorithm 'CLEAN'", *Astron. Astrophys.*, **89**, 377–378.

Cornwell, T. J. (1980), *The Mapping of Radio Sources from Interferometer Data*, Ph. D. Thesis, University of Manchester.

Cornwell, T. J. (1983), "A simple method of stabilizing the clean algorithm", *Astron. Astrophys.*, **121**, 281–285.

Cornwell, T. J. (1984a), "Is Jaynes' maximum entropy principle applicable to image reconstruction?", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 291–296.

Cornwell, T. J. and Evans, K. F. (1985), "A simple maximum entropy deconvolution algorithm", *Astron. Astrophys.*, **143**, 77–83.

Frieden, B. R. (1972), "Restoring with maximum likelihood and maximum entropy", *J. Opt. Soc. Am.*, **62**, 511–518.

Gull, S. F. and Daniell, G. (1978), "Image reconstruction from noisy and incomplete data", *Nature*, **272**, 686–690.

Högbom, J. (1974), "Aperture synthesis with a non-regular distribution of interferometer baselines", *Astrophys. J. Suppl. Ser.*, **15**, 417–426.

Jaynes, E. T. (1982), "The rationale of maximum entropy methods", *Proc. IEEE*, **70**, 939–952.

Narayan, R. and Nityananda, R. (1984), "Maximum entropy—flexibility versus fundamentalism", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 281–290.

Narayan, R. and Nityananda, R. (1986), "Maximum entropy image restoration in astronomy", *Ann. Rev. Astron. Astrophys.*, **24**, 127–170.

Schwab, F. R. (1984b), "Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry", *Astron. J.*, **89**, 1076–1081.

Schwarz, U. J. (1978), "Mathematical-statistical description of the iterative beam removing technique (method CLEAN)", *Astron. Astrophys.*, **65**, 345–356.

Schwarz, U. J. (1979), "The method 'CLEAN'—use, misuse and variations", in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., D. Reidel (Dordrecht, Holland), pp. 261–275.

Segalovitz, A. and Frieden, B. R. (1978), "A 'CLEAN'-type deconvolution algorithm", *Astron. Astrophys.*, **70**, 335–343.

Skilling, J. and Bryan, R. K. (1984), "Maximum entropy image reconstruction: general algorithm", *Mon. Not. Roy. Astr. Soc.*, **211**, 111–124.

Steer D. G., Dewdney, P. E., and Ito, M. R. (1984), "Enhancements to the deconvolution algorithm 'CLEAN'", *Astron. Astrophys.*, **137**, 159–165.

Wernecke, S. J. and D'Addario, L. R. (1976), "Maximum entropy image reconstruction", *IEEE Trans. Computers*, **C-26**, 351–364.

9. SELF-CALIBRATION

Tim Cornwell and Edward B. Fomalont

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

$$\widetilde{V}_{ij}(t) = g_i(t)g_j^*(t)G_{ij}(t)V_{ij}(t) + \varepsilon_{ij}(t) + \epsilon_{ij}(t)\,. \qquad (9\text{–}1)$$

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 9–1 then simplifies to

$$\widetilde{V}_{ij}(t) = g_i(t)g_j^*(t)V_{ij}(t) + \epsilon_{ij}(t)\,. \qquad (9\text{–}2)$$

For simplicity we neglect the effects of time averaging and finite bandwidth, discussed in Lectures 2, 12 and 13; 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 (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

$$\sigma_G^2 = \frac{\sigma_V^2}{(N-3)S^2}\,, \tag{9–3}$$

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 (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 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

[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.

be factored out as two element-related quantities. The loss of absolute source strength information is obvious from Equation 9–2. 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 at least 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, Sec. 2) 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

$$\mathcal{S} = \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) \left| \widetilde{V}_{ij}(t_k) - g_i(t_k) g_j^*(t_k) \widehat{V}_{ij}(t_k) \right|^2 , \tag{9–4}$$

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 9–4 is re-expressed as

$$\mathcal{S} = \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) \left| \widehat{V}_{ij}(t_k) \right|^2 \left| X_{ij}(t_k) - g_i(t_k) g_j^*(t_k) \right|^2 , \tag{9–5}$$

where

$$X_{ij}(t) = \frac{\widetilde{V}_{ij}(t)}{\widehat{V}_{ij}(t)} . \tag{9–6}$$

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,

$$V_{ij,\mathrm{corr}}(t) = \frac{\widetilde{V}_{ij}(t)}{g_i(t) g_j^*(t)} . \tag{9–7}$$

(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

$$\mathcal{H} = -\sum_k I_k \ln \frac{I_k}{M_k e}\,, \tag{9–8}$$

subject to

$$\begin{aligned} \mathcal{S} &= \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) \left| \widetilde{V}_{ij}(t_k) - g_i(t_k) g_j^*(t_k) \widehat{V}_{ij}(t_k) \right|^2 \\ &= \text{expected value}\,, \end{aligned} \tag{9–9}$$

and

$$\sum_k I_k = \text{estimated value of total flux density}\,, \tag{9–10}$$

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 9–2,

$$\widetilde{\phi}_{ij}(t) = \phi_{ij}(t) + \theta_i(t) - \theta_j(t) + \text{noise term}\,, \tag{9–11}$$

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

$$\begin{aligned}\widetilde{C}_{ijk}(t) &= \widetilde{\phi}_{ij}(t) + \widetilde{\phi}_{jk}(t) + \widetilde{\phi}_{ki}(t)\\ &= \phi_{ij}(t) + \phi_{jk}(t) + \phi_{ki}(t) + \text{noise term}\\ &= C_{ijk}(t) + \text{noise term}\,.\end{aligned} \tag{9–12}$$

Thus, for an array of three or more elements, and neglecting noise, closure phase is always a good observable. For an array of 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:

$$\Gamma_{ijkl}(t) = \frac{\left|\widetilde{V}_{ij}(t)\right|\left|\widetilde{V}_{kl}(t)\right|}{\left|\widetilde{V}_{ik}(t)\right|\left|\widetilde{V}_{jl}(t)\right|}\,. \tag{9–13}$$

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 and 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] This terminology is similar to that of closing, or closure, errors in traversed loops, used by surveyors.

(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.

(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. 9–2). 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 and 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 and 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 and 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 16.

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 9–4 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 9–4 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 16 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 $\tau_{\rm sc}$ 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,

$$\sigma_G^2(\tau_{\rm sc}) = \frac{\sigma_V^2(\tau_{\rm sc})}{(N-3)S^2}\,, \tag{9–14}$$

and, for phase correction,

$$\sigma_G^2(\tau_{\rm sc}) = \frac{\sigma_V^2(\tau_{\rm sc})}{(N-2)S^2}\,, \tag{9–15}$$

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:

$$\mathcal{S} = \sum_k \sum_{\substack{i,j \\ i\neq j}} w_{ij}(t_k) \left| \widetilde{V}_{ij}(t_k) - g_i(t_k) g_j^*(t_k) \widehat{V}_{ij}(t_k) \right| \,. \tag{9–16}$$

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, 18 and 20).

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,

$$\mathcal{S} = \sum_k \sum_l \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k, \nu_l) \left| \widetilde{V}_{ij}(t_k, \nu_l) - g_i(t_k) g_j^*(t_k) \widehat{V}_{ij}(t_k, \nu_l) \right|^2 . \qquad (9\text{–}17)$$

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

$$\sigma_{B,C} = \frac{\sigma_{G,C}}{\sqrt{M}}, \tag{9–18}$$

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 (1984b) has outlined a possible solution to this problem, which, owing to its complexity, has never been tried out; similar, but less unwieldy algorithms (C. R. Subrahmanya 1988, in preparation) 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 16.

Bibliography

A good and extensive review article on self-calibration appears in the 1984 Edition of the *Annual Review of Astronomy and Astrophysics* (Pearson and Readhead 1984).

References

Clark, B. G. (1981), "Orders of magnitude of some instrumental effects", VLA Scientific Memorandum No. 137, NRAO.

Cornwell, T. J. (1981), "An error analysis of calibration", VLA Scientific Memorandum No. 135, NRAO.

Cornwell, T. J. and Wilkinson, P. N. (1981), "A new method for making radio maps with unstable radio interferometers", *Mon. Not. Roy. Astron. Soc.*, **196**, 1067–1086.

Cotton, W. D. (1979), "A method of measuring compact structure in radio sources using VLBI observations", *Astron. J.*, **84**, 1122–1128.

Dyson, F. J. (1975), "Photon noise and atmospheric noise in active optical systems", *J. Opt. Soc. Am.*, **65**, 551–558.

Ekers, R. D. (1984), "The almost serendipitous discovery of self-calibration", in *Serendipitous Discoveries in Radio Astronomy*, Proceedings of a Workshop Held at the NRAO on May 4, 5, 6, 1983, K. I. Kellermann and B. Sheets, Eds., NRAO (Green Bank, WV), pp. 154–159.

Hamaker, J. P., O'Sullivan, J. D., and Noordam, J. E. (1977), "Image sharpness, Fourier optics, and redundant spacing interferometry", *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), "A phase sensitive interferometer technique for the measurement of the Fourier transforms of spatial brightness distributions of small angular extent", *Mon. Not. Roy. Astron. Soc.*, **118**, 276–284.

Linfield, R. P. (1986), "A modified hybrid mapping technique for VLBI data", *Astron. J.*, **92**, 213–218.

Muller, R. A. and Buffington, A. (1974), "Real-time correction of atmospherically degraded telescope images through image sharpening", *J. Opt. Soc. Am.*, **64**, 1200–1210.

Pearson, T. J. and Readhead, A. C. S. (1984), "Image-formation by self-calibration in radio astronomy", *Ann. Rev. Astron. Astrophys.*, **22**, 97–130.

Readhead, A. C. S., Walker, R. C., Pearson, T. J., and Cohen, M. H. (1980), "Mapping radio sources with uncalibrated visibility data", *Nature*, **285**, 137–140.

Readhead, A. C. S. and Wilkinson, P. N. (1978), "The mapping of compact sources from VLBI data", *Astrophys. J.*, **223**, 25–36.

Schwab, F. R. (1980b), "Adaptive calibration of radio interferometer data", *Proc. S.P.I.E.*, **231**, 18–25.

Schwab, F. R. (1981), "Robust solutions for antenna gains", VLA Scientific Memorandum No. 136, NRAO.

Schwab, F. R. (1984b), "Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry", *Astron. J.*, **89**, 1076–1081.

Woolf, N. J. (1982), "High resolution imaging from the ground", *Ann. Rev. Astron. Astrophys.*, **20**, 367–398.

10. ERROR RECOGNITION

Ronald D. Ekers

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) provide 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. 10–1a or 10–1d). 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 10–1b, and 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

Figure 10–1. Fourier transform pairs. The broken lines indicate the imaginary domain. For many more examples see Bracewell (1978).

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. 10–2). 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. 10–3a 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 $\mathsf{F}\epsilon$ is added to the image and is independent of the position and amplitude of any other structure in the image, i.e., for which

$$V + \epsilon \rightleftharpoons I + \mathsf{F}\epsilon\,; \tag{10–1}$$

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

$$V\epsilon \rightleftharpoons I * \mathsf{F}\epsilon\,; \tag{10–2}$$

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 10–2. *(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 10–3. 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 10–4. 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

$$V * \epsilon \rightleftharpoons I\,\mathsf{F}\epsilon\,, \tag{10–3}$$

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 6, Sec. 3).

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 (Lectures 2 and 13) 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. 10–4a), this will produce a symmetric error pattern $\mathsf{F}\epsilon$ 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. 10–4b) 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 10–3a and 10–3b.

3. Examples

3.1. Additive errors.

These are errors of the form given by Expression 10–1. 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 13. 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 10–5, 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

Figure 10–5. Top left quadrant of an image of a point source (bottom right corner) with continuous narrow-band interference. [From Thompson (© 1982 IEEE).]

arrays (such as the **D** configuration of the VLA), and it results in an error in the image with a very large angular scale.

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 7. 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 10–2. 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

[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.*

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 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 Lectures 5 and 9. 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 10–3a shows the effect of an amplitude calibration error, and Figure 10–3b 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. 10–2). 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. 10–3a and 10–3b).

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

$$V(u)e^{-2\pi i u\Delta}\,, \tag{10–4}$$

and then by the shift theorem for Fourier transforms we have

$$V(u)e^{-2\pi i u\Delta} \rightleftharpoons I(l-\Delta)\,. \tag{10–5}$$

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

Figure 10–6. Asymmetric pattern which could be caused by atmospheric phase errors.

cause the image to be convolved with an asymmetric error function (see the example in Fig. 10–6). 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 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 9.

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 10–2 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 Lectures 2 and 13. 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 6. The effect of having non-coplanar interferometer baselines and finite computing precision are discussed in Lectures 14 and 16. 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 9.

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:

(i) Possible confusing sources which would either alias into a smaller field or have sidelobes in the smaller field (see Lecture 2) can be recognized.

(ii) Extended emission is more obvious. If unrecognized in a higher-resolution image, it can be mistaken for an error (see Sec. 3.2.1).

(iii) Various checks and computations (e.g., deconvolution) can be performed quickly, because of the smaller size (in pixels) of the low-resolution image.

(iv) 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 16).

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 10–2.

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.

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), "The response of a radio astronomy synthesis array to interfering signals", *IEEE Trans. Antennas Propagat.*, **AP-30**, 450–456.

11. IMAGE ANALYSIS

EDWARD B. FOMALONT

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 9), 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

Figure 11–1. Relative emphases of image features achieved through use of differing resolutions. **(a)** $0''\!.5 \times 0''\!.4$ resolution; **(b)** $2''\!.0 \times 2''\!.0$ resolution; **(c)** $0''\!.5 \times 0''\!.4$ resolution, with high-pass filtering.

of the image can be further modified in order to better discern very small or very large features. Figure 11–1 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).

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

$$I'(l_j) = \frac{1}{N} \sum_i I(l_i) K(l_j - l_i)\,, \tag{11–1}$$

where N is a normalization factor. Extensions to n dimensions are obvious. Some examples of convolution functions are

	$K(\lvert l_i - l_j \rvert)$		
$\lvert i-j \rvert$	(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 11–1 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

$$V'(u_j) = V(u_j)k(u_j)\,. \tag{11–2}$$

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 tasks 'UVMAP' and 'MX'). Even with the extra overhead of two Fourier transforms, many convolutions are more quickly calculated using Equation 11–2. 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', 'MX' 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 11–1, 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 $K(z) = \frac{\sin 2\pi U z}{2\pi U z}$. 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 and 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) differ. There is a slight offset in the pointing axis between the right- and left-hand circular polarizations (in elevation), 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 12, 13, and 14—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

$$\begin{aligned} F &= \sum I(l_i) && \text{Integrated Intensity,} \\ L &= \frac{1}{F}\sum l_i I(l_i) && \text{Mean Position,} \\ B &= \sqrt{\frac{1}{F}\sum (l_i - l)^2 I(l_i)} && \text{Width;} \end{aligned} \qquad (11\text{–}3)$$

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,

$$\sigma^2 = \sum (M(p_j; i) - I(l_i))^2 . \tag{11–4}$$

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 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 11–2 (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 11–3, 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 11–2, 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 13). The component which was fit in Figure 11–2 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 13). 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

Figure 11–2. Image fit to a bright component.

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:

$$\begin{array}{ll} \Delta Z = \Delta R/3 & \text{Zero Bias,} \\ \Delta P = \Delta R & \text{Peak Intensity,} \\ \Delta L = \Delta R W/2P & \text{Position,} \\ \Delta W = \Delta R W/P & \text{Width,} \\ \Delta F = \sqrt{\Delta R^2 + (I\Delta W/W)^2} & \text{Integrated Intensity.} \end{array} \tag{11–5}$$

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

Figure 11–3. 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.

some images which cannot be parametrized, and a suitable display is all that is needed in these cases.

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 11–3 (left). A one-dimensional analysis has been made along lines perpendicular to the jet axis at increasing distances from the core (Killeen, Bicknell, and 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 and 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, and 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

Figure 11–4. 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 11–1.

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 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 11–4.

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 11–3, 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 'MX'). 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 11–4. 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 17 and 18.

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**	$\mathrm{RM} = \frac{\psi(\nu_1)-\psi(\nu_2)}{\lambda_1^2-\lambda_2^2}$
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

Figure 11–5. The noise distribution within an image. *(Left)* A nearly-Gaussian noise distribution. *(Right)* An approximately-Rayleigh noise distribution.

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 11–5. For images on photographic media, the errors scale with the pixel intensity, 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 10 and 24 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 images 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 10).
(2) Different projection of the sky by various instruments (Greisen 1983).
(3) Misalignment of arrays of detectors.
(4) Mosaicing of adjacent images.

The forms of these distortions are generally known, and they can be corrected by using Equation 11–1 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 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 ($\sim 0\overset{''}{.}1$ 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 $0\overset{''}{.}1$ 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), "Nonlinear coordinate systems in AIPS", AIPS Memorandum No. 27, NRAO.

Killeen, N. E. B., Bicknell, G. V., and Ekers, R. D. (1986), "The radio galaxy IC 4296 (PKS 1333−33) I: Multi-frequency VLA observations", *Astrophys. J.*, **302**, 306–336.

Perley, R. A., Bridle, A. H., and Willis, A. G. (1984), "High-resolution VLA observations of the radio jet in NGC 6251", *Astrophys. J. Suppl. Ser.*, **54**, 291–334.

Weast, R. C. and Selby, S. M. (1975), *Handbook of Tables for Mathematics*, Revised Fourth Edition, CRC Press (Boca Raton, FL), p. 865.

12. SPECIAL PROBLEMS IN IMAGING

WILLIAM D. COTTON

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 13, 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 12–1. 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 12–2, 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 $12\overset{\prime}{.}9$ 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 13 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

Figure 12–1. 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.

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 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

Figure 12–2. The effect of bandwidth smearing on a source $12\overset{\prime}{.}9$ northeast of the delay tracking center. The smearing is along the radial direction.

of the source shown in Figure 12–2 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 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, the observing band can be divided

[1] The term *bandwidth synthesis* is frequently used to describe the process of improving the u-v coverage by independently gridding and combining data obtained in several different frequency channels. — *Eds.*

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 13 in terms of the time derivatives of the observing geometry. Due to the complex

Figure 12–3. **(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.

nature of the effect, its symptoms are not as easily recognized as are those due to 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, where the visibility function may be highly variable, 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 distance in the l-m plane 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 12–3 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 24 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 12–4. **(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'.

Figure 12–4 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 12–4a clearly shows the sidelobes of distant confusing sources (one of which is shown in Fig. 12–2). To remove the effects of these distant sources by deconvolving the entire region, a 4096×4096 image would be necessary.

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 12–4. 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 12–4b shows the effect on the image from Figure 12–4a 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 12–4a is 109 μJy and in Figure 12–4b is 62 μJy. It is of interest to note that the bandwidth-smeared image shown in Figure 12–2 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, Sec. 2.3) 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 5). 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

$$V(u,v,w) = \sum_{i=1}^{N} A_i e^{-2\pi i(l_i u + m_i v + n_i w)}\,, \tag{12–1}$$

where

$$A_i = \text{flux density of component } i\,,$$
$$(l_i, m_i) = \text{position of component } i\,,$$
$$\text{and} \quad n_i = \sqrt{1 - l_0^2 - m_0^2} \quad \text{with} \quad (l_0, m_0) = \text{center of the field}\,.$$

The w-term in Equation 12–1 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:

$$\text{error} \approx \pi w \theta^2 \,, \tag{12–2}$$

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 12–2, and the relation "phase error (in turns, i.e., multiples of 2π)" = "position error (radians)" $\times$ "spatial frequency (wavelengths)", the apparent position shift in arcseconds is approximately given by:

$$\text{position error} \approx \frac{\theta^2}{2 \times 2.06 \times 10^5} \sin z \,. \tag{12–3}$$

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 12–3 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 12–5. This Figure shows model source data for a point $47\overset{'}{.}5$ from the

Figure 12–5. **(a)** The response of the VLA to a point model source $47\overset{'}{.}5$ 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.

phase center, for VLA u-v coverage obtained at 40° declination. Figure 12–5a shows the image derived for a full track of the object, and Figure 12–5b shows the image derived for a single 30 minute subset of the data. Figure 12–5a shows a gross distortion of the image as the apparent position of the source changes during the day. Figure 12–5b 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. 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

$$w = au + bv\,, \tag{12–4}$$

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 14.

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 15.

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 12–6.

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

Figure 12–6. 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.

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 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), "Curvature of the sky", VLA Scientific Memorandum No. 107, NRAO.

Clark, B. G. (1980), "An efficient implementation of the algorithm 'CLEAN'", *Astron. Astrophys.*, **89**, 377–378.

Clark, B. G. (1981), "Orders of magnitude of some instrumental effects", VLA Scientific Memorandum No. 137, NRAO.

Clark, B. G. (1982b), "Large field mapping", 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), "Broadband mapping of sources with spatially varying spectral index", VLB Array Memorandum No. 324, NRAO.

Crochiere, R. E. and Rabiner, L. R. (1983), *Multirate Digital Signal Processing*, Chapter 4, Prentice–Hall (Englewood Cliffs, NJ).

Hudson, J. A. (1977), "An analysis of aberrations of the VLA radio synthesis telescope", Chapter 5, Ph. D. Thesis, The American University, Washington, D.C.

Schwab, F. R. (1984b), "Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry", *Astron. J.*, **89**, 1076–1081.

13. WIDE FIELD IMAGING I: BANDWIDTH AND TIME-AVERAGE SMEARING

ALAN H. BRIDLE AND FREDERIC R. SCHWAB

1. INTRODUCTION

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 12—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, as discussed in Lecture 24. We wish now to characterize the two effects mathematically and to justify the approximations embodied in the practical formulae used in the later lecture.

2. BANDWIDTH SMEARING (CHROMATIC ABERRATION)

2.1. 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:

$$I_\nu(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V_\nu(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{13–1}$$

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 6, Section 4.1) that if the functions $X(\mathbf{x})$ and $x(\mathbf{u})$ form a Fourier transform pair in n dimensions, i.e., if $X = \mathsf{F}x$, 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

$$\frac{1}{|\alpha|^n} X\left(\frac{\mathbf{x}}{\alpha}\right) = \mathsf{F}x(\alpha\mathbf{u})\,. \tag{13–2}$$

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 13–1 in terms of the *bandwidth-smeared* intensity $\widetilde{I}(l,m)$ and the *frequency-dependent* u's and v's:

$$\widetilde{I}(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widetilde{V}(u_0,v_0)e^{2\pi i(u_0 l+v_0 m)}\,du_0\,dv_0\,, \tag{13–3}$$

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

$$\widetilde{V}(u_0,v_0) = \frac{1}{\int_{-\infty}^{\infty} G(\nu')\,d\nu'}\int_{-\infty}^{\infty} V\left(u_0\frac{\nu}{\nu_0}, v_0\frac{\nu}{\nu_0}\right)\left(\frac{\nu}{\nu_0}\right)^2 G(\nu')e^{2\pi i\frac{\nu'}{\nu_0}(u_0 l+v_0 m)}\,d\nu'\,. \tag{13–4}$$

Now consider the bandwidth smearing of a point source with unit amplitude. As Equation 13–4 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

$$V(u,v) = e^{-2\pi i u l_0}\,. \tag{13–5}$$

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 6). Inserting the sampled and weighted point source visibility into Equation 13–4,

$$\widetilde{V}(u_0, v_0) = \int_{-\infty}^{\infty} S(u_0, v_0) W(u_0, v_0) e^{-2\pi i u_0 \frac{\nu}{\nu_0} l_0} \left(\frac{\nu}{\nu_0}\right)^2 G_n(\nu') e^{2\pi i \frac{\nu'}{\nu_0}(u_0 l + v_0 m)}\, d\nu' \,, \quad (13\text{–}6)$$

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 13–3 for the bandwidth-smeared intensity can then be written as

$$\widetilde{I}(l, m) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} S(u_0, v_0) W(u_0, v_0) e^{-2\pi i u_0 \frac{\nu}{\nu_0} l_0} G_n(\nu') e^{2\pi i u_0 \frac{\nu'}{\nu_0} l}\, d\nu'\right] e^{2\pi i u_0 l}\, du_0\, \delta(m)\,. \quad (13\text{–}7)$$

Rearranging the exponentials, we can rewrite Equation 13–7 as

$$\widetilde{I}(l, m) = \int_{-\infty}^{\infty} S(u_0, v_0) W(u_0, v_0) e^{2\pi i u_0 (l - l_0)} \left[\int_{-\infty}^{\infty} G_n(\nu') e^{2\pi i u_0 \frac{\nu'}{\nu_0}(l - l_0)} d\nu'\right] du_0\, \delta(m)\,. \quad (13\text{–}8)$$

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

$$d(\tau) = \frac{1}{\int_{-\infty}^{\infty} G(\nu')\, d\nu'} \int_{-\infty}^{\infty} G(\nu') e^{2\pi i \tau \nu'}\, d\nu' \,. \quad (13\text{–}9)$$

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 13–8 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 13–8 can therefore be rewritten, using the convolution theorem, as the convolution of four transforms:

$$\widetilde{I}(l, m) = \mathsf{F}S * \mathsf{F}W * \int_{-\infty}^{\infty} e^{2\pi i u_0 (l - l_0)} du_0\, \delta(m) * \int_{-\infty}^{\infty} d(\tau) e^{2\pi i u_0 l} du_0 \,. \quad (13\text{–}10)$$

The first two convolutions give us the narrow-band image—a "dirty beam" $B = \mathsf{F}S * \mathsf{F}W$ 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 $\mathbf{F}S * \mathbf{F}W * I * D$ *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 13–10, 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—

$$\widetilde{I}(l,m) = \int_0^\infty (B * I)\left(\frac{l\theta_0}{\sqrt{l^2+m^2}}, \frac{m\theta_0}{\sqrt{l^2+m^2}}\right) D(\sqrt{l^2+m^2} - \theta_0, \theta_0)\, d\theta_0\,. \tag{13–10$'$}$$

If $D(\theta, \theta_0)$ had no dependence on θ_0, then Equation 13–10$'$ would reduce to a convolution equation, since the distortion function appearing in the above equation would be a function solely of the difference $\theta - \theta_0$.

2.2. Square bandpass, no tapering, square u-v coverage.

This is the case that was used to illustrate bandwidth smearing in Lecture 12 (see Fig. 12–2).

Passband function:

$$G(\nu') = \begin{cases} 1, & \text{if } |\nu'| < \Delta\nu/2\,, \\ 0, & \text{otherwise}\,, \end{cases} \qquad \text{thus} \qquad \int_{-\infty}^{\infty} G(\nu')\, d\nu' = \Delta\nu\,. \tag{13–11}$$

Delay function:

$$d(\tau) = \operatorname{sinc} \frac{\Delta\nu\,(l_0 u + m_0 v)}{\nu_0}\,. \tag{13–12}$$

Distortion function:

$$D(\theta) = \frac{\nu_0}{\Delta\nu\,\theta_0} \Pi\left(\frac{\nu_0\theta}{\Delta\nu\,\theta_0}\right), \qquad \text{where} \quad \Pi(s) \equiv \begin{cases} 1, & \text{if } |s| < \frac{1}{2}\,, \\ 0, & \text{otherwise}\,. \end{cases} \tag{13–13}$$

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:

$$S(u,v) = \Pi(u/A)\,\Pi(v/A) \tag{13–14}$$

(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:

$$B(l,m) = \operatorname{sinc} Al \operatorname{sinc} Am \tag{13–15}$$

The smeared point source response is B_D, the convolution of the dirty beam B (Eq. 13–15) with the distortion function D (Eq. 13–13). At any given offset $\Delta\theta$ from the beam center at θ_0, the amplitude of this response is

$$\begin{aligned} B_D(\Delta\theta,\theta_0) &= \frac{\nu_0}{\Delta\nu\,\theta_0}\int_{\Delta\theta-\frac{\Delta\nu\,\theta_0}{2\nu_0}}^{\Delta\theta+\frac{\Delta\nu\,\theta_0}{2\nu_0}} \operatorname{sinc} A\theta\, d\theta \\ &= \frac{\nu_0}{\pi A\Delta\nu\,\theta_0}\left[\mathrm{Si}\left(\pi A\left(\Delta\theta+\tfrac{\Delta\nu\,\theta_0}{2\nu_0}\right)\right)-\mathrm{Si}\left(\pi A\left(\Delta\theta-\tfrac{\Delta\nu\,\theta_0}{2\nu_0}\right)\right)\right], \end{aligned} \tag{13–16}$$

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

$$\begin{aligned} \eta &= \pi A\theta_{\mathrm{HPBW}} = 3.79\,, \\ \alpha &= \frac{\Delta\theta}{\theta_{\mathrm{HPBW}}} = \text{offset from peak response in undistorted HPBW's, and} \\ \beta &= \frac{\Delta\nu}{\nu_0}\frac{\theta_0}{\theta_{\mathrm{HPBW}}} = \text{fractional bandwidth} \times \text{radius in HPBW's,} \end{aligned} \tag{13–17}$$

we can rewrite Equation 13–16 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,

$$B_D(\Delta\theta,\theta_0) = \frac{1}{\eta\beta}\left(\mathrm{Si}\,\eta\left(\alpha+\tfrac{\beta}{2}\right)-\mathrm{Si}\,\eta\left(\alpha-\tfrac{\beta}{2}\right)\right). \tag{13–18}$$

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 13–18. 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. 13–18 with $\alpha = 0$ and $\beta = 0$). For this case,

$$R_{\Delta\nu} = \frac{I}{I_0} = \frac{2}{\eta\beta}\,\mathrm{Si}\,\frac{\eta\beta}{2}\,. \tag{13–19}$$

2.3. Square bandpass, circular Gaussian tapering.

For this case, the bandpass, delay and distortion functions are identical to those in the previous example (Eqs. 13–11 through 13–13).

Sampling function:

$$S(u,v)=1 \quad \text{(over an area large relative to the scale of the taper).} \tag{13–20}$$

Weighting (tapering) function:

$$W(u,v)=\frac{\sqrt{\pi}\,\theta_{\mathrm{HPBW}}}{\gamma}\exp\frac{-\pi^2\theta^2_{\mathrm{HPBW}}(u^2+v^2)}{\gamma^2}\,,\quad (\gamma\equiv 2\sqrt{\ln 2}=1.665)\,. \tag{13–21}$$

Dirty beam:

$$B(\theta)=\exp\frac{-\gamma^2\theta^2}{\theta^2_{\mathrm{HPBW}}}\,. \tag{13–22}$$

The calculation of the degraded beam follows that in Section 2.2, leading to

$$B_D(\Delta\theta,\theta_0)=\frac{\sqrt{\pi}}{2\gamma\beta}\left(\operatorname{erf}\gamma\left(\alpha+\tfrac{\beta}{2}\right)-\operatorname{erf}\gamma\left(\alpha-\tfrac{\beta}{2}\right)\right)\,, \tag{13–23}$$

where erf is the usual error function. The reduction in amplitude of a point source relative to zero bandwidth is therefore

$$R_{\Delta\nu}=\frac{I}{I_0}=\frac{\sqrt{\pi}}{\gamma\beta}\operatorname{erf}\frac{\gamma\beta}{2}\,. \tag{13–24}$$

2.4. Gaussian bandpass, circular Gaussian tapering.

Passband function:

$$G(\nu')=\exp\frac{-\gamma^2\nu'^2}{(\Delta\nu)^2}\,,\qquad \text{where}\quad \gamma\equiv 2\sqrt{\ln 2}=1.665\,,$$

$$\text{thus}\quad \int_{-\infty}^{\infty}G(\nu')\,d\nu'=\frac{\sqrt{\pi}\,\Delta\nu}{\gamma}\,. \tag{13–25}$$

Delay function:

$$d(\tau)=\exp\frac{-\pi^2(\Delta\nu)^2(l_0u+m_0v)^2}{\gamma^2\nu_0^2}\,. \tag{13–26}$$

Distortion function:

$$D(\theta)=\frac{\gamma\nu_0}{\sqrt{\pi}\,\Delta\nu\,\theta_0}\exp\frac{-\gamma^2\nu_0^2\theta^2}{(\Delta\nu)^2\theta_0^2}\,. \tag{13–27}$$

For this case, the sampling function is again unity, the weighting function is the Gaussian specified by Equation 13–21, and the dirty beam is the Gaussian specified by Equation 13–22. The degraded beam is

$$B_D(\Delta\theta,\theta_0)=\frac{1}{\sqrt{1+\beta^2}}\exp\frac{-\alpha^2\gamma^2}{1+\beta^2}\,, \tag{13–28}$$

and the reduction in the peak response is

$$R_{\Delta\nu} = \frac{I}{I_0} = \frac{1}{\sqrt{1+\beta^2}} \,. \tag{13–29}$$

2.5. Graphs of the main bandwidth smearing effects.

Figures 13–1 and 13–2, 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 13–9 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, as we shall see in Lecture 24.

3. Time-Average Smearing

3.1. 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 values corresponding to the mid-points t of the averaging intervals, although the data come from time ranges $|\delta t| \le \tau_a/2$ 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 6). 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.

Figure 13–1. The reduction in peak response to a point source, I/I_0, for each of the band shape and taper combinations discussed in Sections 2.2 through 2.4, plotted as a function of the dimensionless parameter β.

Figure 13–2. The broadening of the point source response relative to zero bandwidth, for each of the band shape and taper combinations discussed in Sections 2.2 through 2.4, plotted as a function of the dimensionless parameter β.

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

$$\frac{d\phi}{dt} = 2\pi\nu \left(\frac{du}{dt} l + \frac{dv}{dt} m \right) . \tag{13–30}$$

Averaging a waveform of frequency f for a time τ reduces the response by a factor $\operatorname{sinc} f\tau$, 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

$$R_\tau = \frac{I}{I_0} = \operatorname{sinc} \left(\frac{du}{dt} l + \frac{dv}{dt} m \right) \approx 1 - \frac{\pi^2}{6} \left(\frac{du}{dt} l + \frac{dv}{dt} m \right)^2 , \tag{13–31}$$

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

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} \sin H & \cos H & 0 \\ -\sin\delta\cos H & \sin\delta\sin H & \cos\delta \\ \cos\delta\cos H & -\cos\delta\sin H & \sin\delta \end{pmatrix} \begin{pmatrix} L_X \\ L_Y \\ L_Z \end{pmatrix} , \tag{13–32}$$

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

$$\frac{du}{dt} = \frac{1}{\lambda}\left(L_X \cos H - L_Y \sin H\right)\frac{dH}{dt} \tag{13–33}$$

and

$$\frac{dv}{dt} = \frac{1}{\lambda}\left(L_X \sin\delta \sin H + L_Y \sin\delta \cos H\right)\frac{dH}{dt}\,. \tag{13–34}$$

Substituting these expressions into Equation 13–31 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.

3.2. 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):

$$\left(\frac{du}{dt}\right)^2 = \frac{\omega_e^2}{\lambda^2}\left(L_X^2\cos^2 H + L_Y^2 \sin^2 H - L_X L_Y \sin 2H\right)\,, \tag{13–35}$$

$$\left(\frac{dv}{dt}\right)^2 = \frac{\omega_e^2 \sin^2\delta}{\lambda^2}\left(L_X^2\sin^2 H + L_Y^2 \cos^2 H + L_X L_Y \sin 2H\right)\,, \tag{13–36}$$

and

$$\frac{du}{dt}\frac{dv}{dt} = \frac{\omega_e^2 \sin\delta}{2\lambda^2}\left((L_X^2 - L_Y^2)\sin 2H + 2L_X L_Y \cos 2H\right)\,. \tag{13–37}$$

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

$$\left\langle\left(\frac{du}{dt}\right)^2\right\rangle = \frac{\omega_e^2}{2}\,\frac{L_X^2 + L_Y^2}{\lambda^2}\,, \qquad \left\langle\left(\frac{dv}{dt}\right)^2\right\rangle = \frac{\omega_e^2\sin^2\delta}{2}\,\frac{L_X^2 + L_Y^2}{\lambda^2}\,,$$
$$\text{and} \qquad \left\langle\frac{du}{dt}\frac{dv}{dt}\right\rangle = 0\,, \tag{13–38}$$

from which we get

$$\langle R_\tau\rangle = \frac{I}{I_0} \approx 1 - \frac{\pi^2}{12}\,\omega_e^2\tau_a^2\left(l^2 + m^2\sin^2\delta\right)\frac{L_X^2 + L_Y^2}{\lambda^2}\,. \tag{13–39}$$

Equation 13–39 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

$$\overline{\langle R_\tau \rangle} \approx 1 - \frac{\alpha \pi^2}{12}\, \omega_e^2 \tau_a^2 \left(\frac{\theta}{\theta_{\rm HPBW}} \right)^2 , \qquad (13\text{–}40)$$

which is valid in the regime of small intensity losses. We now evaluate the constant α for a few simple cases:

3.2.1. Square coverage, without tapering.

For square u-v coverage of side A (see Eq. 13–14), the beam is given by Equation 13–15, 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

$$\overline{\langle R_\tau \rangle} = 1 - 1.05 \times 10^{-9} \left(\frac{\theta}{\theta_{\rm HPBW}} \right)^2 \tau_a^2 , \qquad (13\text{–}41)$$

assuming that the loss due to time-average smearing is small.

3.2.2. 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

$$\overline{\langle R_\tau \rangle} = 1 - 1.08 \times 10^{-9} \left(\frac{\theta}{\theta_{\rm HPBW}} \right)^2 \tau_a^2 , \qquad (13\text{–}42)$$

assuming that the loss due to time-average smearing is small.

3.2.3. Circular coverage with Gaussian tapering.

If the array produces a Gaussian beam with FWHM $\theta_{\rm HPBW}$ (Eq. 13–22), the u-v distribution must approximate its transform (Eq. 13–21), so that $\overline{u^2 + v^2} = \gamma^2/\pi^2\theta_{\rm HPBW}^2$ 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

$$\overline{\langle R_\tau \rangle} = 1 - 1.22 \times 10^{-9} \left(\frac{\theta}{\theta_{\rm HPBW}} \right)^2 \tau_a^2 , \qquad (13\text{–}43)$$

again assuming that the loss due to time-average smearing is small.

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), "Orders of magnitude of some instrumental effects", VLA Scientific Memorandum No. 137, NRAO.

Perley, R. A. (1981a), "The effect of bandwidth on the synthesized beam", VLA Scientific Memorandum No. 138, NRAO.

Thompson, A. R. (1973), "The bandwidth effect ('delay beam') for a synthesis array and related requirements for the IF filter characteristics", VLA Electronics Memorandum No. 118, NRAO.

Thompson, A. R. (1982b), "The effects of bandwidth and similar parameters", Lecture No. 5 in *Synthesis Mapping*, A. R. Thompson and L. R. D'Addario, Eds., NRAO (Green Bank, WV).

14. WIDE FIELD IMAGING II: IMAGING WITH NON-COPLANAR ARRAYS

RICHARD A. PERLEY

1. IMAGING IN THREE DIMENSIONS

1.1. The visibility and image volumes.

In Lectures 1 and 2, it was shown that the general response of a two-element interferometer to spatially incoherent radiation from the far field can be written as

$$V(u,v,w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(l,m)e^{-2\pi i\left[ul+vm+w\left(\sqrt{1-l^2-m^2}-1\right)\right]}\frac{dl\,dm}{\sqrt{1-l^2-m^2}}\,. \tag{14–1}$$

This expression is appropriate for a phase-tracking interferometer whose bandwidth $\Delta\nu$ is much less than the observing frequency ν. All symbols in Equation 14–1 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 Equation 14–1 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.

(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 Equation 14–1 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 Equation 14–1 by removing the phase-tracking, i.e., multiply through by $e^{-2\pi i w}$, to give

$$V'(u,v,w) = \iint I(l,m) e^{-2\pi i\left[ul+vm+w\sqrt{1-l^2-m^2}\right]} \frac{dl\,dm}{\sqrt{1-l^2-m^2}} \,. \qquad (14\text{–}2)$$

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 14–2 closely resembles a Fourier relation, however. Indeed, introducing the direction cosine $n = \sqrt{1-l^2-m^2}$ makes Equation 14–2 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

$$F(l,m,n) = \iiint V'(u,v,w) e^{2\pi i(ul+vm+wn)}\, du\, dv\, dw \,, \qquad (14\text{–}3)$$

where V' is defined as in Equation 14–2, 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, if tedious. Substituting Equation 14–2 into Equation 14–3 gives

$$F(l,m,n) = \iiint \left\{ \iint \frac{I(l',m')}{\sqrt{1-l'^2-m'^2}} e^{-2\pi i\left(ul'+vm'+w\sqrt{1-l'^2-m'^2}\right)}\, dl'\, dm' \right\} e^{2\pi i(ul+vm+wn)}\, du\, dv\, dw \,,$$

where I have introduced the dummy variables l' and m' to simplify the notation. This expression can be manipulated to give

$$F(l,m,n) = \iint \left\{ \iiint \frac{I(l',m')}{\sqrt{1-l'^2-m'^2}} e^{-2\pi i u(l'-l)} e^{-2\pi i v(m'-m)} e^{-2\pi i w\left(\sqrt{1-l'^2-m'^2}-n\right)}\, du\, dv\, dw \right\} dl'\, dm' \,.$$

The integrals over u, v, and w can be evaluated using the general result

$$\delta(l'-l) = \int e^{-2\pi i u(l'-l)}\, du \,,$$

to give the expression:

$$F(l,m,n) = \iint \frac{I(l',m')}{\sqrt{1-l'^2-m'^2}} \delta(l'-l)\delta(m'-m)\delta(\sqrt{1-l'^2-m'^2}-n)\, dl'\, dm'.$$

These integrals are elementary, giving

$$F(l,m,n) = \frac{I(l,m)\delta(\sqrt{1-l^2-m^2}-n)}{\sqrt{1-l^2-m^2}}. \tag{14–4}$$

Equations 14–3 and 14–4 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 Equation 14–3. It says that there is a three-dimensional Fourier transform relation between the three-dimensional coherence function $V(l,m,n)$ 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 14–4 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 14–3 and 14–4 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 define 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 6, Section 2.1). The convolution theorem can then be used to show that the result of the Fourier inversion becomes

$$I_d = I * B_d\,, \tag{14–5}$$

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 14–1 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

Figure 14–1. 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.

is shown as a unit circle, and sources of radiation at various values of l are 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 14–1b. 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 Equation 14–5 is a true convolution (if the bandwidth is sufficiently narrow and the effects of time-averaging are small—see Lecture 13), the three-dimensional dirty beam can be deconvolved from the image volume by standard techniques such as 'CLEAN'. Figure 14–1c depicts the result—the "dots" of Figure 14–1a 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 14–1c, where the output image is shown as the projection onto the l-m plane of the 'CLEAN'ed three-dimensional image.

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

Figure 14–2. 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.

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: $V(u, v, w) = e^{2\pi i w} V'(u, v, w)$. 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 14–2a and 14–2b, 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 coordinates (α, δ). For reference, I now give the relations between these,

$$\begin{aligned} l &= \cos\delta \sin\Delta\alpha\,, \\ m &= \sin\delta\cos\delta_0 - \cos\delta\sin\delta_0\cos\Delta\alpha\,, \\ n &= \sin\delta\sin\delta_0 - \cos\delta\cos\delta_0\cos\Delta\alpha\,. \end{aligned}$$

Figure 14–3. 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.

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):

$$\begin{aligned} \delta &= \arcsin\left(m \cos\delta_0 + \sqrt{1 - l^2 - m^2}\,\sin\delta_0\right) , \\ \Delta\alpha &= \arctan\left(\frac{l}{\sqrt{1 - l^2 - m^2}\,\cos\delta_0 - m\sin\delta_0}\right) . \end{aligned} \qquad (14\text{–}6)$$

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 Equation 14–1 is geometric, and has nothing to do with delay. To understand the origin of this term more clearly, refer to Figure 14–3.

Figure 14–3 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 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

$$\phi_{\rm level} = 2\pi ul \,, \tag{14–7}$$

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 $\phi_{\rm ref} = 2\pi w$ 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

$$\phi_{\rm tilt} = 2\pi\left[ul + w\left(\sqrt{1-l^2} - 1\right)\right] \,. \tag{14–8}$$

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 (Lecture 13). What is the effect in three? Lecture 13 showed 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 13, 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 14–4.

The distance from the origin to an object at (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 $\sqrt{l^2+m^2}\,(\Delta\nu/\nu)$.

Figure 14–4. 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.

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 not been tested—they are "paper solutions".

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 14–5. 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

Figure 14–5. 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.

three-dimensional dirty beam. The size of this beam in the n-direction is about the same as in the l- and m-directions, namely: $\theta \approx \lambda/B_{\rm max}$ where $B_{\rm max}$ 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 $\delta n \approx \lambda/2B_{\rm max}$. Figure 14–5 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

$$N_{\rm planes} = n_d/\delta n = B\theta^2/\lambda\,. \tag{14–9}$$

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

$$N_{\rm planes} = \lambda B/D^2\,. \tag{14–10}$$

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

$$N_{\rm planes} = N_l\theta_F/8\,. \tag{14–11}$$

Table 14–1 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 14–1.
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

Above about 2 GHz, the field-of-view to be processed is usually be set by the size of the target object. Background objects whose sidelobes are important can most efficiently be 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]

Furthermore, even at short wavelengths where Equations 14–10 and 14–11 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

[1] As performed, for example, by the AIPS programs 'MX' 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.

[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 .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 14–6. 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.

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 14–6.

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

Figure 14–7. 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.

baseline in wavelengths.) Figure 14–7 shows that the separation 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:

$$\theta_{\rm max} = \sqrt{\lambda/B} \approx \sqrt{\theta_{\rm syn}}\,, \tag{14–12}$$

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 14–7.[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

$$\theta_{\rm max} = \lambda/(2B\theta_0)\,. \tag{14–13}$$

[4] The AIPS multi-field imaging and deconvolution program 'MX' presently operates in this approximation, to minimize computing time and to avoid re-sorting the data.

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 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 14–6 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

$$N_{\rm poly} = 2B\lambda/fD^2\,, \tag{14–14}$$

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 some image computation will be saved through this method. On the other hand, deconvolution by this method will be much 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.

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.

This latter approach may be quite attractive, and will soon be tested at the NRAO.

---

[5] For the VLA in its **A** configuration, this quantity is only $70''$, independent of frequency.

[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.

2.3. Joint deconvolution.

I remarked in the introduction that two-dimensional arrays do not 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 expected.[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 wander on elliptical tracks about the positions given by the standard formulae. It can be shown that the cell positions on the snapshots are given by

$$\begin{aligned} l' &= l + \left(\sqrt{1 - l^2 - m^2} - 1\right) \tan Z \sin\chi \,, \\ m' &= m - \left(\sqrt{1 - l^2 - m^2} - 1\right) \tan Z \cos\chi \,, \end{aligned} \tag{14–15}$$

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. There is no closed-form expression for the inverse of this transformation; 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 $\tan\phi = -\cot\chi$. 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, apparently, has little to recommend it for imaging the entire primary beam at low frequencies. At high frequencies, a given array remains

[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 14–8. 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 14–9. 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.

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.

3. Examples of Three-Dimensional Imaging

Tim Cornwell has written software for 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 14–8 and 14–9 show some examples. Figure 14–8 shows five of the eight planes required in the beam, contoured in the standard fashion. (The 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 14–9 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 14–8. 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.

ACKNOWLEDGMENTS

I thank Robert Braun, Barry Clark, Peter Napier and, especially, Tim Cornwell, for useful discussions on three-dimensional imaging.

15. WIDE FIELD IMAGING III: MOSAICING

TIM CORNWELL

"**Mosaic**, *v. rare*: To combine as if into a mosaic. *Prussia ... is new, and an artificial patchwork, without natural coherence, mosaiced out of bought, stolen, and plundered provinces.* J. L. Motley, *Correspondence*, 18 November 1841."[1]

1. IMAGING A WIDE FIELD OF VIEW

Radio interferometric arrays run into many problems when the desired field of view is large. Among the most important of these are:

- Since the spatial frequency measured by a given pair of elements varies with frequency, the "lens" formed by Fourier transforming the observed visibilities is severely chromatic. Bandwidth smearing of features far from the delay tracking center is inevitable. (See Lecture 13.)
- Earth rotation smears out samples of the visibility function. In the resulting image, features far from the phase tracking center are distorted. (See Lecture 13.)
- For any one direction, the *effective* synthesized aperture results from the projection of a three-dimensional distribution of sample points onto a plane. Hence for a range of directions the synthesized beam varies both in resolution and in the character of the sidelobes. (See Lecture 14.)
- The phase error introduced by propagation of the signal through the atmosphere varies over the field of view. This "anisoplanaticity" necessitates a modified form of self-calibration. (See Lecture 9.)
- The visibility is usually sampled only above some minimum separation $B_{\rm min}$ determined by practical factors such as crosstalk or the desire to avoid collision of the elements! Hence information on structures of size $\gg \lambda/B_{\rm min}$ is lost. (See Lectures 8 and 24.)
- By virtue of its finite diameter D, each element of the array is sensitive only to radiation from a narrow cone of width $\sim \lambda/D$.

In this lecture, I consider the last problem: the limitations imposed by the finite sized field of view of each interferometer element. Progress in this area has been spurred recently by the NRAO's plans to build an interferometric array to operate at millimeter wavelengths. At these short wavelengths, the field of view is comparable to or less than an arcminute, whereas many objects of interest, e.g.,

[1] *Oxford English Dictionary: Condensed Version.*

molecular clouds or nearby galaxies, span many arcminutes. It is therefore vital that we solve this problem. The approach discussed here, based on a Maximum Entropy algorithm, is now in widespread use at the VLA, and at some millimeter telescopes.

Although this limitation may seem to be insurmountable, it is important to remember that single-dish radio telescopes suffer from the same limitation. In that case it is overcome either by pointing the telescope to make total-power measurements for many directions, or by using an array of feeds to collect the power incident from many different directions. Ekers and Rots (1979) noted that by scanning a two-element interferometer in a similar fashion, one could obtain visibility samples over a range of spacings. I discuss this in the next section. In Section 3, I discuss the generalization to include imaging from many different pairs of interferometers, and in Section 4, I describe the current AIPS task, 'VTESS'. Finally, I discuss how this new mode of imaging will affect the design of future interferometric arrays.

2. Scanning a Two-Element Interferometer

A single interferometer (also, somewhat redundantly, called a two-element interferometer) measures the visibility function $V(u,v)$. Under certain conditions (see Lecture 2) this can be written as a Fourier transform of the sky brightness I, multiplied by the antenna primary beam pattern $\mathcal{A}$,

$$V(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}(l,m)I(l,m)e^{-2\pi i(ul+vm)}\,dl\,dm\,. \tag{15–1}$$

The primary beam pattern $\mathcal{A}$ selects just part of the full sky brightness distribution. In the analysis that follows, it will be helpful to label the observed visibility function further with the pointing center (l_0, m_0). Thus

$$V(u,v,l_0,m_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}(l,m,l_0,m_0)I(l,m)e^{-2\pi i(ul+vm)}\,dl\,dm\,. \tag{15–2}$$

This form of the relationship admits the possibility of a pointing center that differs from the phase and delay tracking centers $(l,m) = (0,0)$. If we also assume that the primary beam is independent of position, then the relationship can be written as

$$V(u,v,l_0,m_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}(l-l_0,m-m_0)I(l,m)e^{-2\pi i(ul+vm)}\,dl\,dm\,. \tag{15–3}$$

The virtue of this form is that the concepts of *phase* and *delay tracking* and *antenna pointing* have been separated. Note that the form of this equation is now a Fourier transform of a convolution, which must, of course, be expressible as a product in the Fourier plane.

However, before we get carried away in the mathematics, let us make sure that we understand fully the physical meaning of Equation 15–3. A simple special case occurs for $(u,v) = (0,0)$:

$$V(0,0,l_0,m_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{A}(l-l_0,m-m_0)I(l,m)\,dl\,dm\,. \tag{15–4}$$

Figure 15–1. The range of spatial frequencies measured by a single interferometer of baseline B meters, composed of identical elements having diameter D meters.

This, of course, represents a single-dish observation of the total power coming from the region inside the primary beam. Since the visibility function is just the *correlated* power, the full relationship (Eq. 15–3) states that the correlated power measured by a given interferometer is just the Fourier transform of the region in the primary beam. Therefore, by analogy with the single dish, if we want to image a large region with high angular resolution we must sample many points in the four-dimensional space (u, v, l_0, m_0). So far this is all quite simple—but how do we actually make the image? To answer that, we must follow Ekers and Rots (1979) and continue with the mathematics. Let $i(u, v)$ be the Fourier transform of the brightness distribution $I(l, m)$, and let the conjugate variables to (l_0, m_0) be (u_0, v_0). Further, let $a(u, v)$ be the Fourier transform of the primary beam. Then, using the symmetry of $a(u, v)$, it is easy to show that the Fourier transform of $V(u, v, l_0, m_0)$ with respect to (l_0, m_0) yields

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(u, v, l_0, m_0) e^{-2\pi i(u_0 l_0 + v_0 m_0)}\, dl_0\, dm_0 = a(u_0, v_0) i(u + u_0, v + v_0)\,. \tag{15–5}$$

The physical meaning of this relationship is as follows: suppose that we scan an interferometer in pointing center position (l_0, m_0), and record the measured visibility as a function of the pointing position. A Fourier transform with respect to the pointing position will then yield the Fourier transform of the sky brightness modulated by the Fourier transform of the primary beam. Since the latter extends at most up to spatial frequencies of $\pm D/\lambda$, we can thus recover spatial frequencies spread around the nominal (u, v) by an amount corresponding to a dish diameter.

This result, due to Ekers and Rots (1979), can be understood in simple pictorial terms (Fig. 15–1). A pair of antennas is sensitive to a range of spatial frequencies. In one dimension, the range is from $(B - D)/\lambda$ to $(B + D)/\lambda$ where B/λ is the nominal baseline length. Any single measurement yields the correlated power (i.e., the visibility) for a single linear combination of all these different spatial frequencies. The form of the linear combination is determined by the phase gradient that is introduced across the element apertures by the pointing

Figure 15–2. Information available from two pointings of a one-dimensional interferometer: **(a)** the pointings; **(b)** a_{even}; and **(c)** a_{odd}.

of the elements. By changing the pointing position, we can obtain a different linear combination, and hence, given enough of these linear combinations, we can recover the individual spatial frequencies. The recovery algorithm is simply a Fourier transform.

In practice, scanning is not required. Remember that the Fourier transform of the primary beam must be band-limited (i.e., the antennas must have finite diameter!). The sampling theorem then states that, for full sampling in u-v space, $V(u, v, l_0, m_0)$ is fully specified by its values sampled on a regular grid in (l_0, m_0), where the grid spacing is $\lambda/(2D)$. This makes sense physically: scanning corresponds to looking for changes in the observed visibility function corresponding to minute changes in the pointing position. Since the primary is broad anyway, such fine sampling is unnecessary.

To show that useful information can be derived from just two pointings, consider a one-dimensional example in which the elements of a two-element interferometer are pointed at two positions (see Fig. 15–2). Let V_+ and V_- be the visibility measured at $+l_0$ and $-l_0$. By summing or differencing these we can synthesize different *effective* primary beam patterns, or, equivalently, different aperture patterns. Let

$$V_{\text{even}} = \frac{V_+ + V_-}{2} \quad \text{and} \quad V_{\text{odd}} = \frac{V_+ + V_-}{2i}\,. \tag{15–6}$$

The corresponding effective aperture patterns are

$$a_{\text{even}}(u) = a(u)\cos 2\pi u l_0 \quad \text{and} \quad a_{\text{odd}}(u) = a(u)\sin 2\pi u l_0\,. \tag{15–7}$$

These are drawn in Figure 15–2. Note that these are simply the two possible forms of a Fourier sum for two pointings, as we would expect from the analysis above. By adding more terms, we can increase the "resolution" in the u-plane, and select, with better and better discrimination, just one spatial frequency. One interesting special case occurs when the primary beams are scanned rapidly within one integration time (K. Sowinski, private communication). This produces terms like a_{even}, and so, in one sense, the effective dish diameter is greatly decreased. However, one gives up the ability to separate out spatial frequencies other than B/λ.

To summarize, Ekers and Rots showed that by measuring the visibility from a given interferometer for a range of pointing positions, it is possible to recover the visibility function for a range of spatial frequencies in the neighborhood of the nominal baseline. Although this principle is valid, it is awkward to implement in practice. A different scheme, *mosaicing*, is preferred for observations using interferometric arrays rather than just a simple two-element interferometer. I will discuss this alternative approach next.

3. Mosaicing

Although Ekers and Rots proposed their scheme only for measuring short spacings, it is obvious how to extend the scheme to the imaging of large fields of view. The three necessary steps are:

(1) Scan all interferometers across the required field of view.
(2) Use Equation 15–5 to recover all the spacings.
(3) Make an image of the sky from the full visibility data set, deconvolving if required.

Although this scheme is conceptually simple, it has several drawbacks. First, how can we incorporate knowledge of the noise level? Second, how can we account for limited sampling in the pointing position? Third, the deconvolution step would help in the untangling of the spatial frequencies. It would therefore be advantageous to combine steps (2) and (3).

An alternative approach that overcomes these problems is based on the optimization paradigm for image processing. The philosophy is to form the best image of the sky given all available relevant information. The Maximum Entropy Method (see Lecture 8) is one example of this strategy. In this application, the trick is to regard each measurement of the (fully-labeled) visibility function $V(u, v, l_0, m_0)$ as a constraint on a Maximum Entropy estimation of the sky brightness. We therefore maximize the entropy $\mathcal{H}$,

$$\mathcal{H} = -\sum_k I_k \ln \frac{I_k}{M_k e}\,, \tag{15–8}$$

subject to a constraint on χ^2:

$$\chi^2 = \sum_r \frac{\left|V(u_r, v_r, l_{0r}, m_{0r}) - \widehat{V}(u_r, v_r, l_{0r}, m_{0r})\right|^2}{\sigma^2_{V(u_r, v_r, l_{0r}, m_{0r})}}\,, \tag{15–9}$$

where $\widehat{V}$ denotes the visibility predicted from the model (i.e., from Eq. 15–3).

The crucial quantity in the optimization is the gradient of χ^2 with respect to the ME image I_k, $\nabla\chi^2$. This is used to drive the ME algorithm to the true ME image. In this algorithm, it is calculated in the following ten steps:

(1) Initialize $\nabla\chi^2$ image to zero.

(2) Loop over all visibility samples.
(3) Apply appropriate primary beam pattern to current estimate of ME image.
(4) Fourier transform to obtain predicted visibility $\widehat{V}$.
(5) Accumulate χ^2.
(6) Subtract observed visibility and weight appropriately.
(7) Fourier transform to image plane.
(8) Apply primary beam pattern to result.
(9) Add to $\nabla\chi^2$.
(10) End of loop over visibilities.

Most of these steps are obvious, except for, perhaps, step (8). Although this arises from the mathematical formalism, it also has a simple heuristic explanation. Remember that $\nabla\chi^2$ is used to drive the ME image to the optimum. This step is required to remove the effects of synthesized sidelobes from the residual image; we know that since the elements are insensitive to emission outside the primary beam, any power in the residuals there must originate from sidelobes.

The chief advantages of this scheme are:

- Deconvolution is performed simultaneously with the untangling of the various spatial frequencies. Since the required deconvolution is nonlinear (see Lecture 8), one would expect this to yield results superior to those of separate deconvolution. For example, consider a small object near the overlap point of four pointings. In the joint deconvolution, the object is estimated from the full collection of visibility samples for all four pointings. Compare this to separate deconvolution of each pointing, where the poorer *u-v* coverage of each field will lead to larger errors, even after averaging of all four fields.
- Many different types of data are easily included. For example, differing primary beam patterns are easily accommodated in steps (3) and (8).
- Single-dish observations are simply a special case, and can, in principle, be treated identically to the interferometer measurements. This leads us to a subtly different way of regarding this form of imaging: it is really best described as a hybrid of both single-dish and interferometer observational schemes.
- The weighting of the visibility data is easily maintained.
- The algorithm is well-conditioned: that is, when a particular region is not constrained by observations, then it relaxes to the default image.
- MEM is fast for the extended objects for which this approach is most appropriate.

The chief *disadvantages* are:

- MEM produces resolution that depends on the signal-to-noise ratio, and biased images—see, for example, Narayan and Nityananda (1986). Both effects can be alleviated by convolving the final image with a "clean" beam and adding the residuals (after correction for the primary beams) to obtained a "restored" image similar to that produced by 'CLEAN' (see Sec. 2 of Lecture 8).

- Knowledge of the noise levels and the primary beams is often the limiting factor in the quality of the final image. While information about the primary beam is obviously vital, other deconvolution schemes do not require accurate information about the noise level and may be preferred to the MEM. For example, the Steer–Dewdney–Ito 'CLEAN' algorithm (Steer *et al.* 1984) may be adaptable.

4. 'VTESS'

The algorithm described in Section 3 is implemented in the NRAO's AIPS software package as the task 'VTESS'—Cornwell (1984c, 1988). The ME algorithm used is that developed by Cornwell and Evans (1985); the only major changes required were to modify the routines that calculate χ^2 and $\nabla\chi^2$. The model for the primary beam is that developed by Napier and Rots (1982): a fourth-order polynomial in the square of the distance to the pointing center, truncated at the 7% level. The computations can be arranged so that each iteration requires between 2 and 4 (averaging 3) two-dimensional FFTs per pointing center. The algorithm should converge in about 20 to 30 iterations, making the total CPU time required equal to that required for about 60–90 FFTs per field. Hence the total cost compared to a fully-sampled image, which simply requires one FFT per pointing, is about 60–90 times greater. However, the incremental cost over separate deconvolution is simply in the extra iterations required; usually a factor of two or so.[2]

An example of a mosaiced image of a complex object is given in Figure 15–3*a*. This is a 22-pointing observation of Simeis 57 made with the VLA's **D** configuration at 20-cm wavelength. The pointings are separated by the half-width-to-half-maximum of the VLA primary beam, 15 arcminutes. For comparison, Figure 15–3*b* shows an image of the same object made with the Synthesis Telescope at the Dominion Astrophysical Observatory (DRAO) near Penticton. For the latter, only one pointing was required since the diameter of the DRAO array elements is 8.6 meters. Figure 15–3*c* shows an image formed by linear combination of the separately-deconvolved VLA data sets. The latter is far inferior in its representation of the extended emission, thus verifying the efficacy of joint deconvolution. A detailed investigation (Cornwell 1988) shows that the visibilities estimated from the VLA data agree with those measured by the DRAO telescope on baselines longer than $18\mathrm{m}/\lambda$, whereas the shortest spacing directly measured at the VLA was about $28\mathrm{m}/\lambda$. Therefore, mosaicing has recovered spacings about half a dish diameter shorter than those directly measured. Given that the primary beam model is truncated at twice the beamwidth, this is encouraging. It also underscores the other role for mosaicing: the estimation of

[2] This is not quite accurate since the current version of 'VTESS' is not clever enough to avoid transformation of the blank region outside a given primary beam; this will be fixed, however.

Figure 15–3. **(a)** Mosaiced VLA image of Simeis 57 at a wavelength of 20 cm. The contour levels are 0.3 to 6.0 mJy/pixel, in steps of 0.3 mJy/pixel, where the pixel size is 15 by 15 arcseconds. The tick marks are at intervals of minutes of time in right-ascension and minutes of arc in declination. The crosses refer to the 22 pointing centers used. **(b)** DRAO Synthesis Telescope image of Simeis 57 at 20 cm. **(c)** VLA image of Simeis 57 formed by linear combination of separately deconvolved images.

short spacings for objects that fit into one primary beam. If spacings close to the dish diameter can be measured then mosaicing will recover spacings at about half a dish diameter.

5. Implications for Array Design

Clearly, the possibility of mosaicing will affect the design of some future radio telescopes. There are two major implications: first, since imaging of very wide fields is possible, it becomes important to understand new limiting factors such as far-out primary beam sidelobes produced by the shadow of the feed. Second, some means of measuring very short spacings must be found; for example, a single dish of diameter comparable or equal to the that of the interferometer elements can be used, since, as we saw in the previous section, mosaicing allows estimation of spacings down to about half the dish diameter.

Several factors can limit the quality of the mosaiced image. The most important are:

- Poor knowledge of the primary beam pattern will cause errors. For example, variable sidelobes produced by gravitationally- or thermally-induced warping could be disruptive. The desire for a clean, stable primary beam may drive the antenna design away from conventional alt–azimuth mounted structures with a shadowing feed and feed support legs. Instead, either an equatorial mount or a telescope structure with an off-axis feed geometry may be advantageous.
- Pointing errors can badly corrupt the measured visibilities. The specifications for mosaicing are probably a few times more stringent than usual.

- Interpolation or other errors on bright point sources may dominate. Braun has shown that in some circumstances it is best to deconvolve the high spatial frequencies separately (Braun 1988) since mosaicing is usually of less interest for the fine structure.
- Arranging for consistent calibration between interferometer and single-dish observations has proven to be a thorny problem in the past.
- Bandwidth and time-averaging smearing must both be overcome, as must the non-coplanar baseline effects.

In some circumstances the limiting factor on the speed of mosaicing observations will be the time taken to observe a grid of points spanning an object. The imaging speed for an array that images just one primary beam area goes as ND^2, where N is the number of elements, and D is the element diameter, but if mosaicing is used to image a given solid angle, then the speed goes as ND because although a larger element size will improve the signal-to-noise ratio of a single pointing, many more pointings will be required. This changes considerably the economics of building arrays. The only way around this problem is to use an array of feeds at the focal plane of the interferometer elements. Each beam on the sky would require a separate correlator, but the increase in imaging speed would be tremendous.

6. Concluding Remarks

Although mosaicing as described here is a few years old, much remains to be done. In practical terms, mosaiced observations are not easily processed with current software systems, which were designed for single-pointing interferometric observations. We need to understand the factors that limit the quality of mosaiced images, and we need better diagnostic tools for finding defects in mosaic observations. We also need to investigate alternative schemes for the deconvolution, such as methods not based on the Maximum Entropy Method. Finally, we need to investigate further augmentations: for example, can we self-calibrate pointing errors?

References

Braun, R. (1988), "Mosaicing with high dynamic range", Millimeter Array Memorandum No. 46, NRAO.

Cornwell, T. J. (1984c), "Mosaicing with the mm array", Millimeter Array Memorandum No. 32, NRAO.

Cornwell, T. J. (1988), "Radio-interferometric imaging of very large objects", *Astron. Astrophys.*, **202**, 316–321.

Cornwell, T. J. and Evans, K. F. (1985), "A simple maximum entropy deconvolution algorithm", *Astron. Astrophys.*, **143**, 77–83.

Ekers, R. D. and Rots, A. H. (1979), "Short spacing synthesis from a primary beam scanned interferometer", Proc. IAU Coll. 49, *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., D. Reidel (Dordrecht, Holland), pp. 61–66.

Napier, P. J. and Rots, A. H. (1982), "VLA primary beam parameters", VLA Test Memorandum No. 134, NRAO.

Narayan, R. and Nityananda, R. (1986), "Maximum entropy image restoration in astronomy", *Ann. Rev. Astron. Astrophys.*, **24**, 127–170.

Steer, D. G., Dewdney, P. E., and Ito, M. R. (1984), "Enhancements to the deconvolution algorithm 'CLEAN'", *Astron. Astrophys.*, **137**, 159–165.

16. HIGH DYNAMIC RANGE IMAGING

RICHARD A. PERLEY

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 (which 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

$$V(u) = \delta(u - u_0)e^{-i\phi}\,, \tag{16–1}$$

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

$$I(l) = 2\phi\sin(2\pi u_0 l) + 2\sum_{k=1}^{N(N-1)/2}\cos(2\pi u_k l)\,, \tag{16–2}$$

[1] A snapshot is a single short (< 10 min) observation with a two-dimensional array.

while the beam, or point spread function is

$$B(l) = 2 \sum_{k=1}^{N(N-1)/2} \cos(2\pi u_k l)\,. \tag{16–3}$$

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 10), 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

$$D = \frac{N(N-1)}{\sqrt{2}\,\phi} \approx \frac{N^2}{\sqrt{2}\,\phi}\,, \tag{16–4}$$

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

$$\phi \to \epsilon \quad \text{and} \quad \sin \to \cos\ . \tag{16–5}$$

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

$$D = \frac{\sqrt{N(N-1)}}{\phi} \approx \frac{N}{\phi}\,. \tag{16–6}$$

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

$$D = \sqrt{\frac{N-1}{2}}\,\frac{N}{\phi} \approx \sqrt{\frac{N}{2}}\,\frac{N}{\phi}\,. \tag{16–7}$$

If all antennas have a typical phase error of this magnitude, the dynamic range will decrease by $\sqrt{N}$, yielding

$$D = \frac{1}{\phi}\sqrt{\frac{N(N-1)}{2}}\,. \tag{16–8}$$

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 successive snapshots are independent, then the dynamic range becomes

$$D = \frac{\sqrt{M}\sqrt{N(N-1)}}{\phi} \approx \frac{\sqrt{M}N}{\phi}\,. \tag{16–9}$$

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 200,000 : 1, corresponding to residual errors of 0.001 radians, or 0.06 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 16–7 is justified. From this equation, and inserting the above error, the following limit on the accuracy of the baseline is determined:

$$\delta u = \sqrt{\frac{N}{2}}\,\frac{N}{2\pi D l_0}\,.$$

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 baselines to an accuracy of a few millimeters 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 7, can be written as

$$\delta I = \frac{C}{\sqrt{N(N-1)\Delta\nu\tau}}\,, \tag{16–10}$$

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 200,000 : 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 9 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 1986), 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.

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').

Larry D'Addario 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 and phase, and on the signal powers—and this is the strongest argument for applying the 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, and the information required to apply the corrections off-line is not preserved.

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 16–1 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 $0^\circ\!.03$. (Recall the boxed relation in Sec. 2.)

Table 16–1. 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

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

$$\begin{aligned} v_R &= E_R e^{-i\chi} + D_R E_L e^{i\chi} \\ \text{and} \quad v_L &= E_L e^{i\chi} + D_L E_R e^{-i\chi}\,, \end{aligned} \tag{16–11}$$

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),

$$\begin{aligned} V_{RR} &= V_I(1 + D_{R1}D_{R2}^*) + (V_Q + iV_U)D_{R2}^* e^{-2i\chi} + (V_Q - iV_U)D_{R1}e^{2i\chi} \\ \text{and} \quad V_{LL} &= V_I(1 + D_{L1}D_{L2}^*) + (V_Q + iV_U)D_{L1} e^{-2i\chi} + (V_Q - iV_U)D_{L2}^* e^{2i\chi}\,. \end{aligned} \tag{16–12}$$

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. This prediction is easily tested, and tests to do so will be made soon.

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). An initial attempt to develop AIPS software for correction of all four correlations for the effects of the leakage terms was started but not completed; we still lack this capability.

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

$$R = \frac{Ae^{i\phi_0}}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} e^{i\dot{\phi} t}\, dt\,, \tag{16–13}$$

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

$$\operatorname{sinc} f \Delta t \,, \tag{16–14}$$

where sinc is defined as $\operatorname{sinc} x = (\sin \pi x)/\pi x$, 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

$$\sigma_G = \frac{\sigma_V}{S\sqrt{N}} \,, \tag{16–15}$$

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 Jy), 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.

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 16–2 summarizes the results.

Table 16–2. 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	—	—	—

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 'MX', 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 'MX' for perfect data without phase-shifts (see Sec. 4.9) 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

[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.

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.

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

[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.

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 9, 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 illustrated in Figure 16–1. 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 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

Figure 16–1. 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.

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 Figure 16–2. 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'

Figure 16–2. 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.

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 Figure 16–3. 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 Figure 16–3. It will be immediately apparent that the noise is not uniform. This is generally true for all objects, but espe-

[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.

Figure 16–3. 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.

cially 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 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

Figure 16–4. 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.

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 16–4 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 antenna/IF were then flagged. The subsequent image, shown in Figure 16–5,

[6] Note that the 'UVWORK' file in the AIPS program 'MX' contains these residuals in a form which can efficiently be plotted.

Figure 16–5. 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.

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 Figure 16–3(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 $0^\circ\!.5$ 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.

The procedure is thus as follows:

(1) For the calibrator, apply the best antenna-based calibration, with at least two passes of self-calibration, and careful editing of the residuals.
(2) Divide the u-v data by the transform of the best model (usually represented by a linear combination of a finite number of 'CLEAN' components). It is important that all the flux density present in the visibilities be represented by the model.
(3) Average the quotients, baseline-by-baseline, over the database. Within AIPS, this step is accomplished by the program 'BLCAL'.
(4) Apply the corrections to the source database. Within AIPS, this is done by the program 'SPLIT'.

Some results of applying this procedure to observations of 3C 273 are shown in Figures 16–5 and 16–6. 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

Figure 16–6. 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 16–7. 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 16–8. 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.

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 density. The result of dividing the data by the inverse Fourier transform of the model, as shown in Figure 16–5(left), is shown in Figure 16–6. 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

[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.

than 1%. Application of the mean correlator-based offsets to the data results in the image shown in the center panel of Figure 16–5. 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 Figure 16–5. The dynamic range here is 78,000 : 1.

The dynamic range of the image shown in Figure 16–5 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 16–7 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. At the time of writing (late 1988), this system still does not provide the cross-correlations between the Right and Left channels, so linear polarization is not available. However, this capability should be available in 1989. 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 Figure 16–8a 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 $0\farcs35$, 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 will be dangerous for sources with large spectral index gradients. If the main source of the spectral index difference is in an inverted spectrum core, the problem can be quickly repaired by the same technique as for time-variability—subtraction, from one of the databases, of enough flux density to make the core appear to have the same spectral index as the extended emission.

References

Bagri, D. S. (1986), "Gain residuals in the VLA", Draft.

Clark, B. G. (1978), "Closure—some examples", VLA Electronics Memorandum No. 171, NRAO.

D'Addario, L. R., Thompson, A. R., Schwab, F. R., and Granlund, J. (1984), "Complex cross correlators with three level quantization: design tolerances", *Radio Science*, **19**, 931–945.

Schwab, F. R. (1979), "Quantization correction of VLA correlation measurements", VLA Computer Memorandum No. 150, NRAO.

Thompson, A. R. (1980), "Closure accuracy and related instrumental tolerances", VLA Electronics Memorandum No. 192, NRAO.

17. SPECTRAL LINE IMAGING I: INTRODUCTION

PJOTR ROELFSEMA

1. INTRODUCTION

This lecture discusses some fundamental concepts of spectral line imaging. It does not cover all aspects of spectral line interferometry because some basic ideas are treated in other lectures (e.g., theory in Lecture 1, autocorrelators in Lecture 4, calibration in Lecture 5). Instead, it discusses some practical aspects of making and interpreting spectral line observations.

Section 2 describes the basic properties of radio spectral lines—what they are and what we can, in principle, learn by observing them. It also gives a few examples of spectral line applications. Section 3 describes differences between single-dish and interferometric spectral observations. Section 4 concentrates on the form in which the data from a spectral line experiment finally reach the astronomer: the *data cube.* This section also discusses some aspects of the interpretation and reduction of data cubes to astronomically useful parameters. Lecture 18 gives further details of spectral line analysis.

2. BASIC PROPERTIES OF RADIO SPECTRAL LINES

2.1. What is a spectral line?

This question, although seemingly trivial at first, does not have a simple answer. Even Webster's dictionary does not answer it in a way that immediately rings a bell for a radio astronomer:

> "**spectral line** *n***:** one or a series of linear images of the narrow slit of a spectrograph or similar instrument corresponding to a component of the spectrum of the radiation emitted by a particular source."

The concept of a spectral *line* comes from optical spectra. Historically, light from a bedside candle or the Sun, for example, was projected through a slit onto a prism (see Fig. 17–1). The prism bends the light through an angle that depends on the wavelength. If the light emerging from the prism is projected onto a flat screen, *lines* of different colors are seen, each line corresponding to a given wavelength emitted by the source.

Astronomers could scan a light-sensitive cell along the projection screen, measure the light intensity, and plot this as a function of position along the screen or, its equivalent, of wavelength. Figure 17–2 shows such a plot, but it displays the intensity as a function of frequency (the custom in radio astronomy) rather than of wavelength.

Figure 17–1. The original spectral *line* experiment. Light from a source shines through a slit onto a prism. In the prism the light is dispersed, resulting in a spectrum projected on the screen. The spectral features emitted by the light source show up as lines on the screen.

Figure 17–2 suggests the following definition of an astronomical spectral line:

> "**spectral line** n**:** a high intensity feature in a spectrum, whose extent in frequency space is much less than the central frequency of the feature."

Putting this definition into mathematical terms,

$$S(\nu) = S_c + S_l(\nu, \nu_0, A, \Delta\nu)\,, \quad \text{with} \quad \Delta\nu \ll \nu_0\,. \tag{17–1}$$

Here $S(\nu)$ is the observed spectral energy distribution (in Jy) at frequency ν (Hz), S_c is the continuous component of the spectral energy distribution (in Jy) and S_l is the line component of this distribution (in Jy). As written in Equation 17–1, S_c is independent of frequency in the frequency interval under consideration. Note that this condition may not be met in the real world; S_c may vary slowly with frequency within the observed bandwidth.

S_l is a function that rapidly changes with frequency ν and is characterized by the amplitude A (in Jy), central frequency ν_0 (in Hz) and width in frequency $\Delta\nu$ (in Hz). As an example of such a *line profile* function we have a Gaussian:

$$S_l(\nu) = Ae^{-\gamma^2(\nu-\nu_0)^2/\Delta\nu^2}\,. \tag{17–2}$$

Figure 17–2. The generic spectral line. A feature, narrow in frequency space, is superposed on a background continuum S_c.

In this example, $\gamma = 2\sqrt{\ln 2}$ and $\Delta\nu$ is the full width at half maximum of the Gaussian. The functional form of S_l and the values of the parameters A, ν_0 and $\Delta\nu$ depend on the astrophysical process that governs the emission of the spectral line (see Sec. 2.2). Note that the amplitude A may be negative—in that case we have an *absorption* line.

By measuring A, ν_0, $\Delta\nu$ and, possibly, the shape of S_l the astronomer tries to obtain insight into the physical properties of the material that emitted the spectral line. To do this the astronomer must also estimate the value of S_c—even if it is zero. A line experiment must therefore be designed to observe not only the part of the radio spectrum that contains the spectral line, but also a part of the spectrum with *no* line emission.

2.2. Velocities and the Doppler effect.

Through the Doppler effect, the frequency at which a spectral line is observed can in principle be associated with a velocity of the source of the spectral line emission relative to the observer. This velocity information can be used to study the kinematics of the source.

2.2.1. The general velocity relation.

The general relation between the radial velocity $v_{\rm rad}$, the emitted frequency ν_0, and the observed frequency ν is given by the relativistic expression

$$v_{\rm rad} = c\,\frac{\nu_0^2 - \nu^2}{\nu_0^2 + \nu^2}\,, \qquad (17\text{–}3)$$

where c is the velocity of light. Perhaps because of laziness, astronomers feel that they can get away with approximations to Equation 17–3. Unfortunately many astronomers are also convinced that they themselves are always right while everybody else is wrong, so there are now *two* different "velocity" definitions in common use—a radio and an optical definition.

2.2.2. The radio velocity definition.

The radio velocity definition is obtained by approximating Equation 17–3 for $v_{\rm rad} \ll c$. Thus we obtain

$$v^{\rm r}_{\rm rad} = c\,\frac{\nu_0 - \nu}{\nu_0}\,. \tag{17–4}$$

Here the superscript 'r' in $v^{\rm r}_{\rm rad}$ denotes that this is the radio definition. ν_0 is again the rest frequency of the line, and ν is the observing frequency. This definition is used primarily by astronomers doing galactic radio astronomy.

2.2.3. The optical velocity definition.

To obtain the optical velocity definition, Equation 17–3 is first converted to its equivalent in wavelength by substituting $\nu = c/\lambda$, etc., and then approximating $v_{\rm rad} \ll c$. Thus we obtain

$$v^{\rm o}_{\rm rad} = c\,\frac{\lambda - \lambda_0}{\lambda_0} = cz\,. \tag{17–5}$$

Here the superscript 'o' in $v^{\rm o}_{\rm rad}$ denotes that this is the optical definition. λ_0 and λ denote the rest wavelength of the emitted line and the observing wavelength, respectively. I have also introduced the redshift z. This definition is commonly used in extragalactic radio astronomy. The optical velocity definition in terms of frequency is

$$v^{\rm o}_{\rm rad} = c\,\frac{\nu_0 - \nu}{\nu}\,. \tag{17–6}$$

It must be stressed that *both* the radio and the optical definitions are approximations. The "velocities" derived using these definitions have no physical meaning unless $v_{\rm rad}$ is indeed $\ll c$.

Every astronomer should be aware of these *two* velocity definitions. The velocities derived from Equations 17–4 and 17–5 are *not* interchangeable, as many a newly opened observatory has found out the hard way! For example, when observing the neutral hydrogen (H I) line of a galaxy at a high redshift and using $v^{\rm o}_{\rm rad}$ instead of $v^{\rm r}_{\rm rad}$ to derive the observing frequency using Equation 17–4, the error in the derived frequency can shift the emission from the galaxy out of the observing bandwidth. This happened both at the WSRT and at the VLA in early observations of NGC 1275, a galaxy with an interesting H I absorption feature at $v^{\rm o}_{\rm rad} \approx 8000$ km s^{-1}. As an exercise, try to figure out what total bandwidth should be used to make sure that, despite this error, a feature at 8000 km s^{-1} is still within the observing band.

2.2.4. Velocity rest frames.

For obvious reasons, astronomers are not really interested in the velocity of their favorite source with respect to themselves (the "egocentric" velocity?). Earth rotation makes that velocity vary by 0.5 km s^{-1} over 24 hours and the orbital motion round the Sun makes it vary by 30 km s^{-1} over a year. It is more

Table 17–1. Velocity Rest Frames

Correct for	Amplitude of Correction	Rest Frame Name
Nothing	0 $\mathrm{km\,s^{-1}}$	topocentric
Earth-rotation	$\leq$0.5 $\mathrm{km\,s^{-1}}$	
Earth motion round Earth/Moon barycenter	$\leq$0.013 $\mathrm{km\,s^{-1}}$	geocentric
Earth motion round Sun	$\leq$30 $\mathrm{km\,s^{-1}}$	heliocentric
Solar motion around Solar System barycenter	$\leq$0.012 $\mathrm{km\,s^{-1}}$	barycentric
Solar motion	$\leq$20 $\mathrm{km\,s^{-1}}$	local standard of rest
Galactic rotation	$\leq$300 $\mathrm{km\,s^{-1}}$	galactocentric

useful to define the velocity of a given object with respect to a *velocity rest frame* defined by astronomical objects.

Several such rest frames have been defined, and Table 17–1 lists the main examples. Column 1 gives the motion for which the source velocity is corrected; Column 2 gives the amplitude of the correction; Column 3 gives the name of the rest frame to which the velocity refers after the correction is made.

The choice of rest frame depends on the type of astronomy. For galactic work, the local standard of rest, or LSR, is generally used. For extragalactic work, the barycentric rest frame is generally adopted, but it is usually called "heliocentric". As Table 17–1 shows, this misuse of the name "heliocentric" is of little consequence, because the difference between the heliocentric and barycentric velocity is always less than 0.012 $\mathrm{km\,s^{-1}}$.

2.3. Examples of astrophysically interesting lines.

There are many spectral lines within the frequency domain of modern radio telescopes. As stated earlier, the properties of the observed spectral line depend strongly on the astrophysical process responsible for the emission. When planning spectral line observations, it is therefore important to understand the physical process, and its consequences for line amplitude, line width, etc. I now consider several types of spectral lines in the radio domain to see why we would want to observe them, and how we can observe them.

2.3.1. The neutral hydrogen line.

The neutral hydrogen (or H I) line was predicted by van de Hulst (1945) and was first observed by Ewen and Purcell (1951). The emission process is

Figure 17–3. The emission process for the neutral hydrogen line at 1420.4 MHz.

simple (Fig. 17–3). In a neutral hydrogen atom the spin vectors of the electron and the proton can be either parallel or anti-parallel. When the spin vectors are parallel there is a small, but finite probability that the spin direction of the electron reverses such that the spin vectors are anti-parallel. Since this second state has a lower energy, a photon is emitted in this process, at a frequency of 1420.4 MHz. Although the probability of this transition is so small that it takes any atom $\sim$ 11 million years to change state spontaneously, the number of neutral hydrogen atoms in the universe is so large that the line is readily observable toward many types of object.

As this emission process requires neutral hydrogen to be present, the line is expected from cool neutral material. If a cloud of cool neutral material is located between an observer and a strong source of radio continuum radiation, the transition can be observed in absorption—photons at 1420.4 MHz from the background continuum source are absorbed by atoms in the lower energy state as these atoms make a transition to the higher energy state.

Spiral galaxies (like our own) contain especially large amounts of neutral hydrogen. By observing the neutral hydrogen emission from galaxies, we can study their kinematics and determine their H I content. The intrinsic width of the H I line is a few km s^{-1}, so to study the properties of single H I clouds, a high velocity resolution (e.g., 1 km s^{-1}) should be used to determine the shape of the H I profile. But to study the overall kinematics of a spiral galaxy it is the velocity of the emitting gas at any one point, not the profile shape, that is of interest. It is then not necessary to use a velocity resolution that is high enough to resolve the line shape. A typical spiral galaxy rotates at $\sim$ 200 km s^{-1}, so a velocity resolution $\sim$ 10 km s^{-1} and a total bandwidth equivalent to $\sim$ 600 km s^{-1} (2×200 km s^{-1}, plus a little extra to determine S_c) suffices to determine the velocity field.

2.3.2. Radio recombination lines.

As I hope will be clear from the following description, the term *recombination line* is really a misnomer for this class of radio line. Recombination lines are emitted by ionized plasmas (e.g., H II regions). In such plasmas, electrons and ions are not in general bound to each other. However, occasionally an electron

Figure 17–4. The radio recombination line emission process.

and an ion come so close that the electron is captured, forming an atom or an ion with a lower degree of ionization. When changing from the free state to the bound state the electron emits a continuum photon. Before the atom is again ionized by collisions or by absorbing a high energy photon, the electron will in general cascade down several energy levels. Every such cascade is accompanied by the emission of a line photon at a frequency determined by the difference in energy between the upper and lower levels (see Fig. 17–4). This is called recombination line emission.

A transition in atom X from level $(n + \Delta n)$ to level n is called an X$n\alpha$ transition if $\Delta n = 1$, an X$n\beta$ transition if $\Delta n = 2$, and so forth. The photons emitted in these processes can have essentially every frequency. Recombination lines are indeed observed throughout the radio domain (e.g., H272α at 327 MHz, H76α at 14 GHz) as well as in the infrared (e.g., Brackett α or H4α at 4μm) and optical (Hα or H1α) domains.

These lines are emitted from ionized material, so they offer a way to study the thermal gas in, for example, H II regions, planetary nebulae and ionized low density interstellar matter. As the emission process is relatively sensitive to parameters such as electron temperature $T_{\rm el}$ and density $n_{\rm el}$ of the ionized gas, these lines offer an excellent way to determine these parameters in H II regions. Furthermore, they offer a way to study the kinematics of ionized gas. One advantage of the radio lines as opposed to optical lines such as Hα is that the

Figure 17–5. An illustration of rotational transitions in the formaldehyde molecule. **(a)** rotation projection, and **(b)** the four lowest ortho levels. After Rydbeck and Hjalmarson (1985).

radio emission is not affected by foreground extinction.

Recombination lines are emitted from any region where ionization and recombination occurs. Indeed cold clouds (with T_{el} of a few 100 K) that have an exceptionally low degree of ionization ($\mathrm{N_{H\,II}/N_{H\,I}} \approx 10^{-4}$) also show some radio recombination line emission.

Because an ionized plasma that emits recombination lines will also emit free–free continuum or *bremsstrahlung*, in this case S_c (see Sec. 2.1) is non-zero. In the case of radio recombination line emission from galactic H II regions the continuum is often much stronger than the line (i.e., $S_c \gg A$) making it harder to detect the line.

The intrinsic line width for a radio recombination line from a typical H II region with $T_{\mathrm{el}} \approx 10,000$ K is 10–20 km s^{-1}. Thus, to resolve the line shape a resolution of $\geq$5 km s^{-1} is required. Because turbulence and/or velocity shear in most H II regions is small, a total bandwidth of $\sim$ 100 km s^{-1} is usually enough to observe their hydrogen recombination lines.

2.3.3. Molecular lines.

Many different molecules have been found in the universe, and all emit many different spectral lines. There are two main types of molecular transitions—rotational and vibrational. In a rotational transition, a molecule changes shape due to a rotation of one part of the molecule with respect to the rest of the molecule (see Fig. 17–5). This rotation is accompanied by a change in the energy state of the molecule, and thus by emission of a line photon. Similarly, a molecule may stretch or shrink along one of its axes, thereby changing its energy state and emitting a photon. This is a vibrational transition. The inverse of both processes can occur, leading to absorption lines.

Some molecules that have lines in the radio domain are:

(i) OH: maser emission, first detected in 1963. Used to probe, e.g., circumstellar shells.

(ii) NH_3: observed both in absorption and emission. First detected in 1968. Samples dense cores of molecular clouds. Is used to study material near H II regions, for example.

(iii) H_2O: maser emission, first detected in 1968. Circumstellar envelopes.

(iv) H_2CO: absorption, toward H II regions for example. First observed in 1969. Is used to probe molecular material near H II regions.

There are many more. Many molecular lines are found in the millimeter and sub-millimeter domains and will therefore be interesting prey for millimeter arrays. Molecular material in general is relatively cold, so the lines are narrow, with widths as low as a few tenths of a km s^{-1}. Thus molecular lines must be observed with high (e.g., 0.1 km s^{-1}) velocity resolution.

3. Spectral Line Imaging

Imaging of spectral line data basically means trying to get information about the sky at different frequencies for different positions, or information about the sky brightness distribution at different frequencies. These two different ways to

Figure 17–6. Spectral line imaging with a single dish.

formulate the goals correspond roughly to the ways in which spectral lines are imaged by single dishes or by interferometers.

The result that we aim for in both cases is a three-dimensional matrix $\mathbf{I} = (I_{ijk})$ with intensities I_{ijk} corresponding to the sky brightness distribution $I(x, y, \nu)$ at position (x, y) and frequency ν. Because of its three-dimensional character, $\mathbf{I}$ is usually called a data *cube*. Note that although we call it a cube, the number of elements along the three axes is generally not the same.

In principle we can add polarization and/or time as further variables in $\mathbf{I}$, increasing the dimensionality of the data to 4 or 5. This makes reduction much more difficult but is also bound to result in interesting astronomy. Here I shall only consider the three-dimensional case.

3.1. Spectral line imaging with a single dish.

When observing spectral line emission with a single dish, in principle a single spectrum is obtained, either by using a filter bank, or by using an autocorrelator. By pointing the telescope at a number of positions, spectra can be obtained at each position (see Fig. 17–6). Images can be made by taking, for each position, the signal at a given frequency, and mapping that signal to the location in the data cube or three-dimensional 'matrix' $\mathbf{I} = (I_{ijk})$ corresponding to the position

Figure 17–7. Spectral line imaging with an interferometer.

on the sky:

$$\mathbf{I} = \left\{ I(x_i, y_j, \nu_k) \;\middle|\; i = 1, \ldots, L;\; j = 1, \ldots, M;\; k = 1, \ldots, n \right\}; \tag{17–7}$$

that is, the set of observations consists of complete spectra measured at each of a number of different pointing positions (x_i, y_j). In single-dish observing, one can arrange that the angular coordinates (x, y) are uniformly gridded in right-ascension and declination, or in any other convenient astronomical coordinate system.

3.2. Spectral line imaging with an interferometer.

When observing spectral lines with an interferometer, the signals at a given frequency from a number of antennas are cross-correlated. Next, the obtained visibilities are Fourier transformed to get the sky brightness distribution at that frequency (see Fig. 17–7). Thus we obtain a series of "monochromatic" images of the sky,

$$\mathbf{I} = \left\{ I(l_i, m_j, z_k) \;\middle|\; i = 1, \ldots, L;\; j = 1, \ldots, M;\; k = 1, \ldots, n \right\}; \tag{17–8}$$

i.e, a complete two-dimensional image, in two spatial coordinates, at each of a number of different frequencies ν_i. In general, the l-m angular coordinates are related to astronomical coordinates such as right-ascension and declination, but

Figure 17–8. Obtaining the correlation coefficients for one interferometer pair.

are not uniformly or orthogonally gridded in them (see Sec. 4.1 below). The spectral domain is scanned by using a filter bank, or by Fourier transforming the lag spectrum (see Lecture 4) for each interferometer pair.[1] In both cases we measure, in effect, the visibilities for all interferometers at a given frequency ν_k.

3.2.1. The single interferometer.

Figure 17–8 illustrates the process of spectral line interferometry for one antenna pair as it occurs at the VLA. The signal from two antennas pointed towards the same position in the sky is fed into a cross-correlator. To correct for the delay τ_g of antenna A, this delay is inserted in the signal path of antenna B. A further variable delay τ is added in the signal path of antenna B to obtain the correlation coefficient R at different *lags* τ (see Lecture 4). Thus for coordinates (u, v) in the u-v plane corresponding to the interferometer formed by antennas A and B we measure $R(u, v, \tau)$. By Fourier transforming R with respect to τ we obtain $r(u, v, \nu)$, the correlation coefficient for this interferometer at frequency ν.[2]

Since the lag and frequency domains are related to each other through a Fourier transform, we can determine $r(u, v, \nu)$ with infinite frequency *resolution* only by observing $R(u, v, \tau)$ for all lags between $-\infty$ and ∞. Unfortunately this

[1] This approach is used at the VLA and the WSRT.

[2] After calibration, the correlation coefficient $r(u, v, \nu)$ becomes the *visibility* $V(u, v, \nu)$. Should the distinction between r and R seem unfamiliar, please review Lecture 4.

would mean observing for infinitely long! Thus in practice R is observed only for lags between $-\tau_{\max}$ and $\tau_{\max}$, where $\tau_{\max}$ depends on the desired frequency resolution: $\tau_{\max} = 1/(2\Delta\nu)$. $2n$ lags are used, but (in the case of the VLA correlator) since one is a zero lag, a perfectly valid lag is set to zero to regain symmetry.

Furthermore, R is determined only at a few different lags between $-\tau_{\max}$ and $\tau_{\max}$, because to observe all lags would again require an infinite amount of time. The lag resolution is $\Delta\tau = 1/(2B)$, where B denotes the bandwidth. Measuring R at 'infinite' lag resolution would in principle allow us to measure the spectrum from $\nu = 0$ to $\nu = \infty$, but this would also require an instrument that is about equally sensitive to all frequencies! Thus in practice $R(u, v, \tau)$ is measured at only $2n$ lags at $\Delta\tau$ intervals:

$$\left\{ R(u, v, i\Delta\tau) \,\middle|\, i = -n, \ldots, n \right\}. \tag{17–9}$$

The lag spectrum is unlike the autocorrelation spectrum from a single telescope; the latter is real and even, so only positive lags are required. For the cross-correlator, the lag spectrum is real, but has no symmetry; so both positive and negative lags are needed.

Sampling out to only $\pm n\Delta\tau$ and at $\Delta\tau$ intervals has some nasty consequences. The sampling limit at $\pm n\Delta\tau$ causes the final spectrum to be convolved with a $\operatorname{sinc} x$ function with a FWHM of $1/(2n\Delta\tau)$, resulting in ripples across the band (the Gibbs ripple, see Lectures 4 and 18). The sampling also aliases the spectrum, but this is not a problem if the signal is filtered so that only a total bandwidth $B < 1/\Delta\tau$ is observed.

3.3. Calibration.

The practical differences between the calibration of spectral line and continuum data from an interferometer are small because two assumptions are made:

(1) that there are no correlator-dependent errors in the data;
(2) that the antenna gains, and thus the correlator gains, can be decomposed into two components—one depending *only* on the central observing frequency ν_0 and time, the other *only* on frequency in the bandpass.

The gain for an interferometer pair i, j at a frequency ν_n can then be written in terms of amplitude and phase as

$$A_{ij}(t, \nu_n)\, e^{i\phi_{ij}(t,\nu)} = b_i(\nu_n)\, b_j(\nu_n)\, e^{-i(\beta_i(\nu_n) - \beta_j(\nu_n))} A_{ij}(t, \nu_0)\, e^{i\phi_{ij}(t,\nu_0)}. \tag{17–10}$$

Here $b_i(\nu_n)$ and $\beta_i(\nu_n)$ are the amplitude and phase of antenna i at the frequency ν_n corresponding to channel n, with respect to its amplitude and phase at frequency ν_0.

With these assumptions, line data can be calibrated in two steps. First, we calibrate the time-dependent $A_{ij}(t, \nu_0)$ and $\phi_{ij}(t, \nu_0)$ using data from a set of line channels that are summed to obtain a pseudo-continuum channel ('channel 0') that has a higher signal-to-noise ratio than an individual line channel.

Figure 17–9. An empty data cube.

Next, we calibrate the shape of the bandpass $b_i(\nu_n)b_j(\nu_n)$ and the phase term $\beta_i(\nu_n) - \beta_j(\nu_n)$ for each interferometer assuming they are constant throughout the observation. To do this, we observe a strong source and compare the visibilities measured in channel n with those measured in channel 0.

4. Data Cubes

4.1. What do they look like?

Section 3 introduced the form of spectral line data after calibration and transformation—the three-dimensional data "cube" $\mathbf{I}$, containing samples of our best estimates of the sky brightness $I(l, m, z)$. A typical line observation may have 64 or more channels in the z-direction. Channel images of 256 by 256 pixels (in l and m) are also typical. Thus a typical data cube may consist of more than $256 \times 256 \times 64$, or 4.2×10^6 pixels! To satisfy the sampling criterion, each resolution element should contain at least 2 pixels, so the number of independent resolution elements in the data cube is $\sim 500,000$ pixels. The resolution in the three-dimensional cube is of course characterized by three half-widths: in l, in

Figure 17–10. A data cube with continuum emission.

m and in z; roughly, in R.A. (e.g., arcsec), Declination (arcsec) and velocity (in $\mathrm{km\,s^{-1}}$).

The final result, after analysis by the astronomer, may for example be a rotation curve from H I data with only about twenty numbers. Thus the spectral line astronomer really does data *reduction*! However, to make such data reduction possible the astronomer must understand the behavior of the signal in the data cube. The first job is therefore to find the signal!

Figure 17–9 shows what a data cube looks like. It is a structure with three axes l, m and z containing the intensity as function of these three coordinates. In general the l- and m- coordinates are associated with right-ascension and declination. The z-coordinate is usually associated with frequency or radial velocity (see Sec. 2.1). It should be noted that the pixels in this cube are equidistant in l and m but *not* in right-ascension and declination. Lines of constant right-ascension or constant declination are curves through the cube, which for small angular fields of view are close to straight lines parallel to l and m respectively. Also the pixels are equidistant in z and not in velocity. However, in most obser-

Figure 17–11. A data cube with line emission.

vations the pixels along the z-axis *are* equidistant in frequency. In Figure 17–9, nothing was detected, so the cube is empty.

4.1.1. A data cube with continuum emission.

Given the relatively narrow total bandwidth that is used in typical line experiments, the intensity of the continuum emission can be regarded as almost independent of frequency. In a three-dimensional data cube this implies $I(l, m, z) = I(l, m, 0)$ for all z. The signature of continuum emission in a line data cube is that intensity is independent of channel number. In the cube, a continuum source can be seen as a structure that runs from the front to the back of the cube without changing shape (see Fig. 17–10).

4.1.2. A data cube with line emission.

Line emission differs from continuum emission in that its intensity *does* depend on frequency: $I(l, m, z) \neq I(l, m, 0)$. The shape of a line source therefore changes with frequency (Fig. 17–11). The signature of line emission in the cube is a cloud of high intensity points distributed along an axis that is *not* in general parallel to the z-axis.

Spectral line data will in general contain a combination of line and continuum emission, so that the intensity has two components—$I_l(l,m,z)$ and $I_c(l,m,z)$, where $I_l(l,m,z) \neq I_l(l,m,0)$ and $I_c(l,m,z) = I_c(l,m,0)$.

4.1.3. Finding the signal in a data cube.

Finding the signal in a data cube is not straightforward. When analyzing line data, you are trying to see correlated signal in a three-dimensional structure. A noise spike should show up in the data cube as something with the size of the three-dimensional beam. The source under investigation could be a point source, in all three coordinates. Only statistics can then tell how significant the corresponding spike in the cube is, and the astronomer must use the statistics to judge whether or not to believe it. Signal from an extended source is extended in one or more of the three cube-coordinates, so finding and analyzing extended emission requires more than simple statistics. There is no simple recipe for analyzing extended emission. It requires some feeling for how the emission is distributed through the data cube, and for what information must be extracted from the cube.

In general the signal is explored first by *looking* carefully at the data cube in various ways. For different kinematical properties of the observed source, different—we hope, recognizable—patterns should be present. Examples of different ways to look at the data are:

(1) Look at a "movie" of the data, i.e., display the channel images one after each other at a high speed, so that signal can be seen to 'move' through the cube.

(2) Look at different planes in the cube; e.g., l-m plane (or image plane), l-z or m-z (position-velocity) planes. A rotating galaxy (see below), for example, has a clear signature in the position-velocity plane along its major axis. Also, make movies in different planes.

(3) Look at a three-dimensional display of the data (this is sophisticated image processing that may require much computation and special-purpose display hardware). Use a constant intensity level to define the location of the signal, and display this three-dimensional contour as a solid body.

The next step, after having recognized the signal, is to generate scientifically interesting parameters from it. Again the method depends strongly on the purpose of the observation. We must often derive at least the line integral, and a mean velocity or velocity of peak intensity, as well as the line width for every l-m pixel. There are many different ways to derive these parameters. Some of them are good, and some of them have pitfalls depending on what is to be achieved. Here I shall comment on a few of the methods; more detailed descriptions of most of these are given in Lecture 18 and by Bosma (1978).

(1) Fit a preconceived profile shape for the spectrum at every pixel in the l-m plane. This requires, and thus biases the results toward, some *a priori* knowledge of the profile shape. For example, when

observing radio recombination line emission from H II regions, the spectra are usually well described by a Gaussian, and fitting Gaussians is a reliable way to determine line parameters. However, H I line emission from a rotating galaxy often shows skewed profiles and fitting a Gaussian can be a bad way to estimate the rotation velocity at a given position in the galaxy.

(2) "Blotch" (i.e., mask) regions with no signal by looking at the data in the l-m, l-z or m-z planes, then determine moments with respect to the z-coordinate: the line integral $\mathcal{I} = \int I(v)\,dv$, the intensity weighted mean velocity $\langle v \rangle = \int vI(v)\,dv/\mathcal{I}$ and the dispersion $\sigma^2 = \int v^2 I(v)\,dv/\mathcal{I}$. Because the parameters $\langle v \rangle$ and σ are weighted by the intensity, these parameters are extremely sensitive to noise (see Bosma, 1978).

(3) "Blotch" regions with no signal on the basis of, for example, data smoothed to lower resolution (and thus higher sensitivity to extended structure). Then proceed as in (2).

These three methods can be applied in any plane of the data cube. One way to estimate how well they perform is to try them all and compare the results. Still, it is in this phase that the worst part of the telescope comes in to play—the astronomer's mind. So beware of "interesting" features in a velocity field—check whether you can get rid of them by re-analyzing the data in a different fashion. If they stick around, try to find their signature in the original channel images. What is present in the final product should also be there in the raw data. On the other hand, if a possibly interesting 4σ peak does not pop up right away in, e.g., the integrated line intensity, try to get it out by re-analyzing too!

4.2. Examples of data cubes.

To illustrate some points made above, I shall now discuss two different data cubes—the cube for a uniformly expanding shell (as might be found in circumstellar maser emission) and that for a typical H I observation of a spiral galaxy. The aim here is to get a feeling for the signature of such kinematic structures in the data cube.

4.2.1. An expanding shell.

A uniformly expanding shell is easy to understand in terms of its behavior in the data cube. Such a shell expands away from its center with velocity $\mathbf{v}_{\rm exp}$ (a vector) with an amplitude $\mathbf{v}_{\rm exp}$ that is independent of the direction of $\mathbf{v}_{\rm exp}$. Thus the velocity $v_{\rm rad}$ along the line of sight to such a shell can be calculated easily for each part of the shell as it is projected on the sky. For an infinitely thin shell with a radius R, the radial velocity at a distance r away from the center of expansion is

$$v_{\rm rad}(r) = \pm|\mathbf{v}_{\rm exp}|\sqrt{1-\frac{r^2}{R^2}}\,. \qquad (17\text{–}11)$$

The $+$ sign holds for points on the shell at the far side of the center, the $-$ sign for points at the near side. The loci of constant $v_{\rm rad}$ are circles with radii $r(v_{\rm rad})$ as given by Equation 17–11. When observing line emission from such a shell, we

Figure 17–12. A uniformly expanding shell as seen in an emission line using an interferometer, in theory.

can expect to see rings of emission with different radii in the various channels of our data, as in Figure 17–12. Figure 17–13 shows data for such a case: the 1612-MHz OH maser emission from a circumstellar shell in OH 127.8−0.0 (Booth *et al.* 1981).

Note that an expanding shell will appear as an ellipse of emission in the l-z or m-z position-velocity planes.

4.2.2. H I in a spiral galaxy.

The behavior of the H I emission in a data cube as observed towards a rotating galaxy is more complicated. In the simplest case the neutral hydrogen

Figure 17–13. A uniformly expanding shell as seen in an emission line using an interferometer, in practice. The four channel images show the spatial distribution of the integrated OH emission in the velocity intervals marked a to d in the integrated spectrum (upper panel). After Booth *et al.* (1981).

(and stars and what have you) in a spiral galaxy rotate around its center at a velocity $v_{\rm rot}$ which depends only on the radius r to the center. And the rotation curves are often simple; in the inner part of the galaxy they rise at a constant rate, in the outer part the rotation velocity is constant (see Fig. 17–14c).

If we view such a galaxy face-on, the rotation has no component along the line of sight, so we see all the H I at the same velocity. When the galaxy is inclined to the line of sight, different parts are seen at different velocities. It takes some time to get the projections right for each case, but if the data

Figure 17–14. **(a)** surface density of neutral hydrogen in a typical spiral galaxy. **(b)** model iso-velocity contours for a spiral galaxy inclined to the line of sight with a rotation curve as shown in panel **(c)**.

are analyzed properly the iso-velocity contours for such a galaxy form so called "spider" diagram (Fig. 17–14*b*). Those of you who want to try to derive the formulae and to generate a model velocity field (it is instructive!) can check your answers in Begeman (1988). As we observe the line emission at a set of velocities, we expect to see the shape of the iso-velocity contours in the channel images (Figs. 17–15 and 17–16).

Along the major axis of the galaxy, no projection occurs other than that resulting from the inclination i of the galaxy. Along this we should therefore see

Figure 17–15. A rotating spiral galaxy as seen in H I emission using an interferometer, in theory.

Figure 17–16. A rotating spiral galaxy as seen in H I emission using an interferometer, in practice; UGC 2885.

Figure 17–17. **(a)** major axis position-velocity diagram for UGC 2885. Superposed on the H I contours are the rotation velocities as observed in optical emission lines. **(b)** rotation curve for UGC 2885 as derived from the H I data shown in Figure 17–16.

the rotation curve decreased by $\sin i$. Indeed, if an l-v (position–velocity) diagram is constructed along the major axis, the H I emission follows the rotation curve (Figs. 17–15 and 17–17).

References

Begeman, K. G. (1988), *HI Rotation Curves of Spiral Galaxies*, Ph. D. Thesis, University of Groningen.

Booth, R. S., Kus, A. J., Norris, R. P., and Porter, N. D. (1981), "Observations of a circumstellar shell around the OH/IR star OH127.8−0.0", *Nature*, **290**, 382–384.

Bosma, A. (1978), *The Distribution and Kinematics of Neutral Hydrogen in Spiral Galaxies of Various Morphological Types*, Ph. D. Thesis, University of Groningen.

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

Ewen, H. I. and Purcell, E. M. (1951), "Radiation from galactic hydrogen at 1,420 Mc./sec.", *Nature*, **168**, p. 356.

Rydbeck, O. E. H. and Hjalmarson, Å. (1985), "Radio observations of interstellar molecules", in *Molecular Astrophysics*, G. H. F. Diercksen, W. F. Huebner and P. W. Langhoff, Eds., NATO ASI Series C Vol. 157, D. Reidel (Dordrecht, Holland), pp. 45–175.

van de Hulst, H. C. (1945), "Radio waves in space", *Ned. Tijdschr. Natuurk.*, **11**, 201.

Webster's Third New International Dictionary of the English Language (1971), G. & C. Merriam Company (Springfield, MA).

18. SPECTRAL LINE IMAGING II: CALIBRATION AND ANALYSIS

JACQUELINE H. VAN GORKOM AND RONALD D. EKERS

1. INTRODUCTION

This lecture discusses problems specific to the calibration and analysis of multi-frequency aperture synthesis observations. Such observations may be taken either to measure the kinematic or physical conditions in line-emitting regions (as discussed in Lecture 17) or in continuum studies when the monochromatic approximation is inadequate (see, for example, Lecture 12). Spectral synthesis differs from continuum synthesis in more ways than just the number of frequencies. The differences will be discussed here.

2. BANDPASS CORRECTIONS

Astronomical calibration of the complex antenna-based bandpass function is needed. Fortunately, the standing wave modulation of the bandpass—which is a large and most uncertain error in single dish observations—is not a problem, because the receiver noise is not correlated between elements.

Usually the bandpass is calibrated by observing a calibrator long enough to reach the required signal-to-noise ratio. Bandpass normalization by the auto-correlation spectrum for each antenna can be used to correct for the amplitude variations across the band, but this does not correct for the phase variations. If there is likely to be line emission or absorption toward the calibrator (e.g., as in galactic H I observations), then one can calibrate the bandpass by shifting the frequency enough to avoid the line—e.g., by shifting symmetrically by plus and minus a few MHz and averaging the two observations.

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*, that must be considered. The Gibbs phenomenon is the ringing around sharp changes in the frequency spectrum that occurs because the temporal cross-correlation measurements are truncated. The Gibbs phenomenon has more serious consequences in spectral line imaging than in an autocorrelation spectrometer. For the autocorrelation spectrometer, the correlation function must be real and even (so that no negative lags need be measured), because the power spectrum is known to be real and even. However, the cross-correlation function, though real, is not even, and so it yields a spectrum of complex source visibilities. As a result we will find the mirror image/complex conjugate of the "real" spectrum at negative frequencies. This causes a phase discontinuity at $\nu = 0$ if the observed visibility phase is nonzero (see Lecture 4). Since the effect of the Gibbs phenomenon now depends on the visibility phase, the frequency ripple will change with position in the image and with a change in the instrumental phase. The only place where the effect will be correctly calibrated by the bandpass calibration is at the position of the bandpass calibrator (provided that the instrumental phase has not changed between bandpass calibration and source observation). The effect can be attenuated by a suitable tapering of the cross-correlation measurements before the Fourier transform (e.g., by Hanning smoothing), but this significantly degrades the frequency resolution.

To determine the optimum bandpass calibration method it is useful to consider three categories of observations:

(1) *Line source without a continuum background.* If the line emission is confined to the flat part of the receiver bandpass, then the resulting spectrum has no discontinuity at the bandpass edges. This is the only case where it may be advisable to use different cross-correlation tapers for the source and for the calibrator. A uniform taper can be used for the source spectra to maximize the frequency resolution while Hanning smoothing should be used for the calibrator. The passband edges will then be calibrated improperly.

(2) *Line source and one continuum point source.* Here, the Gibbs effect can be calibrated out completely, provided that the relative positions of the point source and calibrator coincide (e.g., if both are at the field center). The same weighting must be used for source and calibrator—it may be either uniform or Hanning.

(3) *Line source and extended continuum emission.* The Gibbs effect will be different for the source and calibrator and cannot be calibrated out. Hanning smoothing is highly recommended for both the calibrator and the source.

Finally, here are a few other general comments on the Gibbs effect. The peak amplitude of the ringing does not decrease as the number of lags, or, equivalently, the number of frequency channels, is increased. However, as the number of lags is increased, the effect becomes more confined to the neighborhood of the discontinuity. In particular, as the number of channels is increased, the Gibbs effect produced by the bandpass skirts becomes more concentrated at the band edges. Because the Gibbs phenomenon can be calculated for the observed structure, it could, in principle, be included in a three-dimensional deconvolution procedure. For an extensive discussion of the Gibbs effect, see Bos (1984, 1985).

3. Chromatic Aberration

As discussed in Lectures 2, 12 and 13, synthesis radio telescopes have an inherent chromatic aberration because they form an image by adjusting the phase $\Delta\phi$ of the correlated signals from each point in the image plane instead of the arrival time of the wavefront $\Delta t = \Delta\phi/2\pi\nu$ at each point. This aberration causes a radial smearing that increases linearly away from the delay tracking center—the point in the image for which the time delays were equalized by the delay tracking system. It is commonly known as *delay beam smearing.*

To overcome this defect it is necessary to subdivide the band into channels that are narrow enough to have negligible aberration, and to use the actual frequencies of each channel rather than the band center frequency when constructing the images. In most spectral line observations the channel bandwidth criterion is automatically satisfied by the spectral resolution requirements, but for broad-band continuum observations using a spectral line correlator the number of channels required will be determined by the smearing.

Even if delay beam smearing is negligible in the individual spectral line channels, the sidelobe structure corresponding to the array geometry still changes with frequency. This must be taken into account when analyzing the spectral line data cube. The effects of frequency-dependent sidelobes may be reduced by deconvolution procedures, but it is preferable to proceed as far as possible without introducing the additional deconvolution uncertainties, discussed in Section 7. Consider two extreme cases:

(1) If the spectral line images are computed using the correct frequencies for each channel, then the structure in each image will all be in the correct place, but the sidelobe pattern will change from channel to channel by a radial scaling factor proportional to frequency. This is preferable if the spectral line images contain much weak, frequency-dependent structure spread over a large field.

(2) If the same frequency is (artificially) used for all channels, then the sidelobe patterns will all be the same, but the structure in the image will now be distorted by a radial scaling factor proportional to the actual frequency offset. (If all these channels were added, this would give the delay beam smearing.) If the dominant source is near the field center, the sidelobes will be correctly removed in channel differences and the image distortion may be acceptable.

4. High Spectral Dynamic Range

By *spectral dynamic range* we mean the channel to channel stability or, in other words, the ratio between the continuum and the residual that is left when two line-free channels are subtracted from each other. Some spectral line observations require high spectral dynamic range. For example, in a recombination line observation the line to continuum ratio may only be a few percent. The spectral dynamic range must be better than 1000 : 1 at every point in the image if the line is to be measured to 10% accuracy. Fortunately, the phase errors caused by the atmosphere—a primary source of error in continuum observations—are independent of frequency and can be removed by subtracting a continuum image formed by averaging frequency channels outside the line emission. The spectral dynamic range is limited by multiplicative errors (see Lecture 10) that differ from one frequency channel to the next because bandpass calibration is imperfect. The effect of baseline-dependent errors has not been investigated, but it is believed to cause problems at a level well below residual antenna-based errors (see Lecture 16). The Fourier transform of the error term is convolved with the strong continuum source distribution, so errors that differ from one channel to the next remain when the continuum is subtracted. Examples are bandpass calibration errors in amplitude and phase. The patterns that arise from gain

Figure 18–1. An example of the effects of small phase errors on difference images. **(a)** shows images without phase error; **(b)** shows a similar observation, but with a phase error that increased with distance from the band center, and changed sign at the center. As a consequence, the line emission seems to move from right to left with increasing velocity. **(c)** shows twice a line image of the first observation. At left, one telescope has a phase error of 0.5 degree, at right the two baselines with that telescope have been removed.
The contour intervals are: for the continuum, (a) 500, (b) 770 mJy/beam; for the line, (a) 15, (b) 25, (c) 10 mJy/beam.

and phase errors are well known, but they may still be hard to recognize, since they are convolved with extended source structure. An additional problem is that if the errors vary systematically with frequency then it may be impossible to distinguish velocity structure in the line emission from chromatic errors. Figure 18–1 shows two examples of the effects of small phase errors on difference images, using weak recombination line emission from the compact H II region DR 21. In one example a small frequency-dependent phase error causes an apparent (and erroneous) rotation of the H II region. The second example shows that a phase error of only half a degree on one baseline produces sidelobes in a line-minus-continuum image that are almost as strong as the line. The errors can most easily be found by differencing two (preferably line-free) channel images. A powerful way to locate errors is to look at the Fourier transform of a difference image.

5. Continuum Subtraction

Before subtracting the continuum image from the line channels, the possible effects of the chromatic aberration discussed in Section 3 must be considered. The easiest way to subtract a continuum is to form an image from the average of frequency channels on either side of the line emission and subtract that from all channels—before doing any deconvolution. Because of the frequency structure of the sidelobes, this method is satisfactory when the total bandwidth $\Delta\nu$ is small compared to the observing frequency, e.g., $\Delta\nu/\nu_0 \ll 0.01$, and the continuum emission is not too strong. If the frequency dependence of the sidelobe structure is important, then it may be necessary to subtract the inverse Fourier transform of a model of the continuum from the measurements in the visibility plane, using the exact frequencies and geometry to calculate the u-v coordinates of the visibility samples. This model might consist of the positions and amplitudes of discrete continuum sources. This is convenient if there are many confusing continuum sources in the field but the line emission is confined to a small region. One can then make one large image of the line-free channels, find the continuum sources, subtract these in the u-v plane from all channels—and then make a set of small images to study the line emission. Alternatively, a model could be derived using the deconvolution procedures discussed below. This is especially useful when the continuum source is strong. Figure 18–2 shows an extreme example—weak hydrogen emission from the strong (200 Jy) radio galaxy Centaurus A. In this example, the mean continuum image was 'CLEAN'ed and the 'CLEAN' components were subtracted from all frequency channels in the u-v plane to remove most of the strong continuum correctly. Finally, the mean remaining continuum was imaged and subtracted from all frequency channels. The noise in the resulting line-minus-continuum images is a factor ten lower than in an image obtained by simply subtracting the continuum image from a line channel. The example illustrates how a line can remain undetected if the continuum is subtracted improperly.

6. Self-Calibration

When applying the self-calibration relations to correct for spectral line ob-

Figure 18–2. An example of the effects of chromatic aberrations on a line-minus-continuum image. **(a)** shows the continuum source, Cen A. The contours range from 1 to 18 Jy/beam. **(b)** shows a line-minus-continuum image at 260 km/s. The dirty mean continuum image has been subtracted from the dirty channel image. **(c)** same as (b), but now the continuum has been subtracted by taking the 'CLEAN' components of the continuum and subtracting these in the *u*-*v* plane from the channel image. **(d)** as (c) after 'CLEAN', the H I is real!
The contour intervals are: (b) 260, (c) and (d) 26 mJy/beam. The total flux of the continuum source is 200 Jy, which is why the residuals in (b) are so high that the line remains undetected.

servations for atmospheric phase errors, the same set of phase corrections can be used at all frequencies. The best way to determine these corrections depends on the nature of the image. If there is strong continuum, then the average of all continuum channels can be used to obtain the self-calibration solution. However, if the line emission is much stronger than the continuum, then the channel with the highest signal-to-noise ratio in the line could be used (e.g., a channel containing a maser line). If all the channels have similar intensity and spatial structure changing with frequency it may be necessary to develop a three-dimensional model to be used for a global self-calibration solution (Ekers and van Gorkom, 1984)

7. Deconvolution

The deconvolution of spectral line images presents some special problems. Consider the case of an absorption line experiment involving strong and complex continuum emission. If individual channels are deconvolved independently then a small fractional error in the deconvolution can substantially increase the errors when these channels are subtracted to form the line-minus-continuum image. Figure 18–3 shows an example of this with the 'CLEAN' algorithm—the H I absorption in the radio galaxy NGC 315. 'CLEAN'ing the continuum and line images separately and then subtracting the continuum image produces spurious emission along the jet. If the channels are subtracted before deconvolution this problem is avoided, because the strong continuum signal and its sidelobes are removed and do not have to be deconvolved. If, on the other hand, the line channels contain less continuum signal—for example, because there is strong absorption—then it makes sense to do the subtraction or other image manipulation after the deconvolution. The image that results after subtraction can contain a mixture of positive and negative features, so algorithms that rely on positivity to suppress sidelobes (e.g., MEM) cannot be used.

We often want to measure the ratio of a deconvolved line-minus-continuum to a deconvolved continuum image, to determine the distribution of optical depth or electron temperature (recombination lines). For this purpose we have an additional constraint on the deconvolution algorithm: it must produce estimates of the real sky as seen through identical transfer functions (i.e., with the same beam).

Finally, information in adjacent channels may be needed to optimize the deconvolution. For example, if uniform weighting is applied in the transform from lag to frequency domain, then the sidelobes in the frequency domain may be deeper than the spatial sidelobes. A three-dimensional deconvolution algorithm would then be essential.

8. Profile Analysis

The optimization of the signal-to-noise ratio in many spectral line problems is equivalent to a matched filter problem where the filter characteristics are determined by the data to be filtered (adaptive filtering). Consider the problem of using a set of spectral line images to determine the integrated properties of the line emission at each point in a rotating object. Such properties would be the integrated emission, its velocity and velocity dispersion (the zeroth, first, and second moments of the velocity profile).

That is, we visualize the set of spectral line images as a set of profiles at various positions in the sky (grid points) and calculate the zeroth, first, and second profile moments. Let $I_i(\alpha, \delta)$ be the surface brightness at velocity v_i, Δv the velocity separation between channels, and $n_{\rm tot}$ the total number of channels; then (in the case of an H I observation) the hydrogen column density is given by

$$N_{\rm H\,I}(\alpha, \delta) \propto \Delta\nu \sum_{i=1}^{n_{\rm tot}} I_i(\alpha, \delta)\,, \tag{18–1}$$

the intensity weighted mean velocity by

$$\overline{v} = \frac{\sum_{i=1}^{n_{\rm tot}} I_i(\alpha, \delta) v_i}{\sum_{i=1}^{n_{\rm tot}} I_i(\alpha, \delta)}\,, \tag{18–2}$$

Figure 18–3. An example to illustrate how 'CLEAN' can introduce spurious emission or absorption features. **(a)** shows the continuum source NGC 315. **(b)** shows an image at the peak of the H I absorption line. The peak of the core is 400 mJy/beam; in (b) it is reduced to 250 mJy/beam. Both images have been 'CLEAN'ed, and that process has, erroneously, slightly increased the surface brightness of the jet in the absorption image. As a consequence, the optical depth image in **(c)** shows negative optical depth along the jet. **(d)** shows the difference image of continuum and the peak absorption channel. The dirty images have been subtracted, and the difference has been 'CLEAN'ed. No pseudo-emission can be seen.
Contours: in (a) and (b) 15 to 250 mJy/beam; (c) letters are negative, contours 0.09; (d) 5 mJy/beam.

and the velocity dispersion by

$$\sqrt{\langle v^2 \rangle} = \sqrt{\frac{\sum_{i=1}^{n_{\rm tot}} I_i(\alpha,\delta)(v_i - \overline{v})^2}{\sum_{i=1}^{n_{\rm tot}} I_i(\alpha,\delta)}}\,. \tag{18–3}$$

The result is to collapse the frequency dimension and produce one image for each of the moments. Coming back to the rotating object, in each channel the emission appears at a different position because of the Doppler shift. One obvious way to get the total emission is to take the sum over all channels. However, at any point in the image the line emission is present in only a few channels—the rest contain noise—so this straightforward procedure increases the noise

Figure 18–4. An example of four images of the H I emission in NGC 1097. The velocity is shown in the upper left corner. The contour interval is 6 mJy/beam.

level significantly over that in individual channels. As an example, Figure 18–4 shows the neutral hydrogen emission in four velocity channels of the barred spiral NGC 1097. The total hydrogen determined by summing all the channels is shown in Figure 18–5. The increase in noise is clear. Several methods have been developed to avoid this signal-to-noise degradation by including only the channels that contain line emission.

Methods that have been used to separate the line signal from the noise in the profiles are:

(1) Apply an *acceptance gate* in intensity ('CUTOFF' method),
(2) Apply an acceptance gate in velocity ('WINDOW' method),
(3) Study each profile on a graphics display of an interactive computer,
(4) Fit the line profile to a preconceived shape,
(5) First smooth the data, then use either (1) or (2).

Methods (1) and (2) have been compared by Bosma (1981), and methods (3) and (4) by van der Kruit and Shostak (1982).

8.1. 'CUTOFF' method.

This method was first described by Rogstad and Shostak (1971). To exclude points with no line emission, a cutoff in surface brightness is applied. For example, all points with $I_i(\alpha, \delta)$ smaller than twice the r.m.s. noise σ are set to

Figure 18–5. A total hydrogen image of NGC 1097. The image has been obtained by taking the sum of all channel images. The contour interval is 210 mJy km/s.

zero and thus do not contribute to the calculation of the above quantities. As an example, the total hydrogen image of NGC 1097 using a cutoff of 2σ is shown in Figure 18–6. Although the image is considerably improved, the method has one serious disadvantage. The calculated moments are subject to systematic effects depending on the cutoff value used. If there are weak features just below the cutoff value, then the calculated zeroth moment (Eq. 18–1) will be too small. The calculated radial velocity will be biased toward the middle of the velocity range, because of noise peaks above the cutoff value. If the line emission is at extreme velocities, then the noise peaks are likely to be at less extreme velocities, thus moving the mean value toward less extreme velocities (see Fig. 18–7). The velocity dispersion can be biased either way. The cutoff of weak wings on the profiles can make it too small, while noise peaks above the cutoff value can considerably increase the calculated velocity dispersion, because of the strong influence of the factor $(v_i - \overline{v})^2$ in Equation 18–3.

8.2. 'WINDOW' method.

This method has been developed to overcome the biases introduced by the 'CUTOFF' method (Bosma, 1978 and 1981). Around each profile a window or acceptance gate in velocity is chosen. In this way the influence of noise peaks outside the H I emission range is eliminated (Fig. 18–7). In practice, the velocity of the peak of the profile, v_0, is determined first, a narrow window is centered around it, and the mean intensity I_{mean} of the points outside the window is computed. The window is then made larger and larger, and each time I_{mean} is calculated. Obviously the value I_{mean} should converge (to zero if no continuum emission is present), and the iteration is stopped if two later values of I_{mean} differ by less than a specified value—e.g., by less than σ/n_c where n_c is the number of channels outside the window. This convergence criterion has been determined experimentally by Bosma (1981).

Figure 18–6. The same as Figure 18–5, except that points below a 2σ cutoff were not included in the sum.

Figure 18–7. Some examples of WSRT profiles observed in the galaxy NGC 5033. N denotes the noise level (1σ), C and V are the results for continuum and velocity from the window method, while C' and V' are the results from the cutoff method. The profile integrals are shown with coarse shading. The fine shading shows the additional contribution to the integral from the cutoff method. The cutoff level used is 1.5σ (Bosma, 1978).

8.3. Interactive study of individual profiles.

This is a variation on the 'WINDOW' method, in which each profile is inspected on a display and the window is set interactively using a graphics input device. Although often superior, this introduces the possibility of personal bias, and is time consuming for large databases.

8.4. Fit the line profile to a preconceived shape.

This method first locates at each position on the sky the highest point in the profile and then fits a Gaussian (or Voigt) profile to the data around those points. Fitting a baseline (continuum) can easily be included. No bias is introduced by a cutoff, but one may be introduced if the assumed functional form is incorrect.

Figure 18–8. An illustration of the use of different smoothing functions to determine the acceptance window. **(a)** spatial smoothing with twice the resolution, **(b)** Hanning smoothing, **(c)** Hanning smoothing over 5 channels, **(d)** spatial + Hanning smoothing.

Some limited testing has shown that this method and 'WINDOW' give similar results for the zeroth and first moments (van der Kruit and Shostak, 1982), while the velocity dispersion determined by fitting is far more reliable.

8.5. Hybrid method.

An improvement to both the 'CUTOFF' and the 'WINDOW' methods is obtained by using data that are smoothed either spatially or in velocity to determine the points to be included, then going back to the full-resolution data for the moment calculations. Figure 18–8 shows examples of this method in which different smoothing functions have been applied. Note that different smoothing functions bring out the hydrogen in different parts of the galaxy.

Acknowledgments

We thank Pat Palmer and Arnold Rots for comments on the manuscript.

References

Bos, A. (1984), "On ghost source mechanisms in spectral line synthesis observations with digital spectrometers", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 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. (1981), "21-cm line studies of spiral galaxies. I. Observations of the galaxies NGC 5033, 3198, 5055, 2861 and 7731", *Astron. J.*, **86**, 1791–1824.

Ekers, R. D. and van Gorkom, J. H. (1984), "Spectral line imaging with aperture synthesis radio telescopes", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 21–32.

Rogstad, D. H. and Shostak, G. S. (1971), "Aperture synthesis study of neutral hydrogen in the galaxy M101", *Astron. Astrophys.*, **13**, 99–115.

van der Kruit, P. C. and Shostak, G. S. (1982), "Studies of nearly face-on spiral galaxies. I. The velocity dispersion of the H I gas in NGC 3938", *Astron. Astrophys.*, **105**, 351–358.

19. VERY LONG BASELINE INTERFEROMETRY I: PRINCIPLES AND PRACTICE

R. CRAIG WALKER

1. INTRODUCTION

Very Long Baseline Interferometry (VLBI) is the technique that lets us use widely separated antennas as elements of an interferometer array. It is distinguished from other forms of radio interferometry by not requiring communication between the elements while the observations are being made. Independent, high-quality frequency standards replace a distributed local oscillator signal, and the base-band, digitized signals are recorded on magnetic tape for later correlation. Satellite links have been used for test experiments, but most VLBI observations are made with separate frequency standards and tape-recorded data.

Since no communication is required between the array elements, the baselines can be arbitrarily long, but ground-based experiments are limited by the size of the Earth to baselines of about 11,000 km. Preliminary observations have been made using NASA's TDRSS satellites to extend the longest baselines to about two Earth radii. These studies show that space VLBI is practical, and dedicated VLBI satellites are now being planned. The typical resolution achieved in ground-based experiments is about 1 milli-arcsecond at 6 cm wavelength, and scales inversely with wavelength. VLBI observations at 1.3 cm are now common, and 3 mm observations have been made. To visualize typical VLBI resolving powers, note that 1 milli-arcsecond corresponds to about 1 a.u. at one kiloparsec (a typical galactic distance) and about 1 pc for an object with a red shift of 0.067 ($H_0 = 100\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$). It is roughly the angular diameter of an orange in Los Angeles as seen from Europe.

There are no fundamental differences between VLBI and the connected-element interferometry discussed in earlier lectures. However, some operational differences complicate the data analysis and make VLBI calibration techniques differ in detail from those used for connected-element interferometers. The most important differences stem from the use of separate frequency standards (requiring that frequency and time offsets be determined from the data), and from the difficulty in determining the geometry of the array (including atmospheric and ionospheric effects) accurately enough to allow the use of phase calibration sources. Traditional VLBI methods also differ from those of connected-element interferometry for historical reasons or because of differences in the types of sources normally observed. The use of tape-recorded signals and delayed correlation does not introduce significant differences. This lecture concentrates on areas where VLBI differs from connected-element interferometry.

2. VLBI Systems

The VLBI technique was developed in the late 1960s to study the bright sources in the nuclei of active galaxies and quasars and to study masers in our own galaxy. The antennas belonged to existing observatories and were not usually designed for use as part of an interferometer. Two recording systems were developed in 1967: the Mk I system in the United States recorded digital data on standard computer tapes, while the system developed in Canada recorded analog data on video tapes. After a few years, the Mk II system was developed in the United States and is still in use world-wide. It records 4 Megabits per second (Mbs) of digital data on video tapes. Three generations of video tape recorders have been used with the Mk II system. The current version uses home-style video cassettes and records four hours of data per tape. There are Mk II correlators at the NRAO in Charlottesville (3 stations), Caltech (16 stations), Bonn (3 stations), Italy (5 stations, the old Caltech correlator), and in Shanghai, China. Other cassette-based systems are being developed or operated in Canada and the Soviet Union.

The Mk III system was developed more recently, primarily at the Haystack Observatory. It uses instrumentation tape recorders and (in its standard wide-band mode) records 28 separate 4 Mbs signals at once on a single tape, for a total bit rate of 112 Mbs. In this mode, each tape can record 13 minutes of data. The system is being upgraded with narrow heads that allow twelve passes over a single tape for about 3 hours of recording. The Very Long Baseline Array (VLBA)—currently under construction by the NRAO—uses an enhanced version of the Mk III system that can record at least 128 Mbs for 12 hours per tape. There are Mk III processors capable of processing 4 or 5 stations at Haystack, the Naval Observatory, Caltech, Bonn, and in Japan. The processors at the Naval Observatory and in Japan are used primarily for geodesy.

The frequency standards used for VLBI are either rubidium vapor frequency standards or hydrogen masers. For the relevant time scales (a few seconds to a few hours) hydrogen masers are the best available standards. Unfortunately, their quality is reflected in a cost of several hundred thousand dollars per maser. On short time scales, crystal oscillators are better, and the VLBI local oscillator signals are generally derived from a crystal oscillator locked to a maser. It is ironic that the best standards for absolute time, the cesium beam standards, do not work for VLBI. They are better for time measurement than the rubidium or maser standards because their absolute frequencies are less sensitive to environmental factors; but if the environment is carefully controlled, rubidium and maser standards provide a more stable signal.

The antennas used as the elements of VLBI arrays are still mostly ones that are used primarily for single-dish observations. To understand what this means for VLBI observations, imagine trying to use a connected-element interferometer if all its antennas were of different designs, worked to different maximum frequencies, had sensitivities ranging over a factor > 100, and were operated and scheduled by different organizations, many with primary responsibilities to other user communities![1] You should then understand some reasons for building the VLBA, and be able to anticipate some problems that arise when calibrating VLBI data.

3. Data Flow

This section reviews the data flow for typical VLBI observations, emphasizing those operations that differ from normal methods of connected-element interferometry. Sections 4 to 6 discuss some aspects of the data reduction in detail, and Lecture 20 discusses the special problems of spectroscopic VLBI.

3.1. Data acquisition.

The signals are received and amplified at each antenna using normal radio-astronomical equipment. They are mixed with a local oscillator signal derived from the VLBI frequency standard, to bring one edge of the observing passband to 0 Hz. Current practice is to use the same total local oscillator frequency at all stations, but the frequencies can be offset to remove partially the Doppler shifts caused by the Earth's rotation, or to avoid zero fringe rates. The mixing and amplification generally occur in several intermingled steps.

The signals are then filtered and digitized. The quantization is to two levels in all current systems but will be to two or four levels on the VLBA. The sampling is typically at the Nyquist rate, the reciprocal of twice the bandwidth. The VLBA will be able to oversample to improve the signal-to-noise ratio for narrow band (usually spectral line) data. (Lecture 4 discussed the effects of digitization on the signal-to-noise ratio.) After digitization, the data are formatted and recorded on magnetic tape. Extra information, notably the time, is encoded with the data. The tapes are shipped to one of several correlators within days of the observations.

3.2. Correlation.

The tapes are correlated on specially-built digital equipment. Each correlator has a maximum number of stations that it can process simultaneously. Modern experiments typically involve more than ten stations so, on most existing correlators, it is necessary to make many passes to process all baselines. The VLBA correlator is designed to handle up to 20 stations for the most common kinds of experiment, and most experiments should be processed in one pass (the VLBA itself will have 10 antennas). The 16 station Mk II capability on the Block II processor at Caltech can also process in one pass large experiments for which the Mk II system has enough bandwidth.

The correlator aligns the tapes in time; reads and decodes the data; aligns the data streams from each station, accounting for clock offsets and geometric delays; shifts the frequency of one data stream from each baseline to account for clock rate offsets and the differential Doppler shift between the stations arising from Earth rotation; multiplies each pair of data streams for a range of delays to generate a correlation function; accumulates the results; and writes the results to an archive medium, usually a magnetic tape. For spectral line data

[1] Imagine also that the "instrument" is allowed to observe for a few weeks at a time and is then dismantled and reassembled before the next observing run! — *Eds.*

(for all data on the VLBA), the correlator may transform the correlation function to a spectrum and may generate an autocorrelation function for each data stream. The frequency shift that removes clock rate offsets and Doppler shifts is accomplished by multiplying one of the data streams for each baseline by a digital approximation (usually 3-level) of the desired sine wave. Two shifted data streams are formed, using sine waves that are 90° out of phase. Both streams are correlated with the data from the other antenna of the baseline to obtain the complex components of the correlation function. This dual fringe rotator is one of the several ways to obtain a complex correlation function that were discussed in Lecture 4.

The VLBA correlator will use the FX architecture discussed in Lecture 4. This creates some differences from the above sequence. The tape alignment and delay removal, to the nearest bit, is the same as above, but a complex, many-level fringe rotator is used to remove the clock rate offsets and Doppler shifts for each station without the losses associated with the 3-level rotators used in current correlators. The individual data streams are then Fourier transformed to form spectra that are to be cross correlated. Before cross correlation, a phase slope is applied across the spectrum to effectively set the delay to a small fraction of the sample interval. This avoids the losses and other problems associated with the inability of XF correlators to set the delay to better than one sample period. After cross correlation, the spectra are similar to those generated by Fourier transforming the correlation function from traditional correlators (see Lecture 4 for a discussion of the subtle differences).

3.3. Editing and fringe fitting.

The data are edited after correlation but before imaging. Any large blocks of bad data should be deleted before the fringe fitting (discussed below), but at this stage the volume of data can be huge (typically 2-second records in current practice), and thorough editing may be impractical. Badly discrepant points or unusually short integrations should be eliminated, but a few low amplitude points will not seriously affect fringe fitting. After fringe fitting, the data are averaged and the data volume is more manageable. At that point, it is worth inspecting all the data and eliminating any bad points. Interactive programs that use a TV or graphics on a CRT are available for data editing.[2]

After correlation, continuum data are *fringe fitted.* This operation, which is critical to VLBI, solves for, and removes, any residual delay and fringe rate offsets. Such offsets arise from a combination of uncertainties in *a priori* clock parameters and in the geometry and atmosphere. Fringe fitting is not needed for connected-element interferometers because the *a priori* uncertainties in delay and rate can easily be made insignificant.[3] Section 4 discusses fringe fitting in detail. It is hoped that the geometry and clocks for the VLBA will be understood

[2] For example, the task 'TVFLG' in AIPS or the program 'IED' in the Caltech VLBI package.

[3] For connected-element interferometers, as long as the instrumental delays are determined at least once each time an antenna is moved, the errors in the *a priori* delays will usually be small relative to the spacing of the lags in the correlation function. Only the zero-delay correlation need be determined for continuum data, and there is no uncertainty in the zero lag of the correlation function used for spectral line observations. The phase is far more sensitive but can

well enough that fringe fitting—and even phase calibration—can be done on calibration sources, to allow imaging of weak program sources. Such observations can be made now, but only with great effort.

3.4. Calibration.

After fringe fitting, the data resemble those that would be obtained from a connected-element interferometer whose phases are uncalibrated. The amplitudes can be calibrated in several ways, though it is traditional to start by using system temperatures and gain curves rather than interpolated calibrator observations. This is partly because there are few, if any, unresolved calibrators for VLBI baselines, and most compact sources are variable. VLBI phase calibration depends on self-calibration (Lecture 9), sometimes referred to in the VLBI literature as *hybrid mapping*. The procedure was originated in the context of VLBI by Readhead and Wilkinson (1978) and by Cotton (1979) and was extended to include amplitude calibration by Readhead *et al.* (1980).

Methods to gauge convergence of self-calibration in VLBI differ somewhat from those used for arrays such as the VLA, partly because of the smaller amounts of data in typical VLBI observations, and partly for historical reasons. I suspect that both communities can learn from each other in this area. Once images are made, there is little difference between VLBI data and other interferometer data, except that in VLBI absolute position information is lost by the fringe fitting/self-calibration process whereas in connected-element interferometry an initial external calibration can often be applied accurately enough to be useful.

3.5. Spectral line data.

In spectral line VLBI, the program source data cannot usually be fringe fitted. Delay offsets that would be removed by the fringe fitting show up as phase slopes across the frequency band in spectral line data. There are two ways to deal with such phase slopes: (a) observe and fringe fit a nearby continuum source to determine the offsets, or (b) locate the features with methods that use only the rate of change of phase (fringe rate) and not the phase itself. Both approaches are discussed in Lecture 20. Typical VLBI spectral line sources (usually masers) have bright features that can be used as phase calibrators, allowing traditional imaging methods (not dependent on self-calibration) to be used on most of the spectral channels. Bright spectral lines also allow use of autocorrelation spectra to calibrate the amplitudes—not self-calibration in the usual sense, but almost as good. Once the data are calibrated, the analysis and display problems resemble those of other radio spectral line data. One difficulty of many VLBI line experiments stems from the nature of the sources. Masers often have compact individual features that can be studied with milli-arcsecond resolution, but the features are spread over several arcseconds of sky. If brute-force methods are used, prohibitively large image sizes are needed; some ways around this are discussed in Lecture 20.

---

be calculated accurately enough that any error is relatively constant with time and can usually be calibrated using observations of known sources (normal phase calibration). Certainly the residual fringe rate is very small and does not limit the integration time.

4. Fringe Fitting

For any interferometer, the correlated signal appears at a delay and phase that depend on the geometry of the interferometer, the source position, the atmosphere and ionosphere, and on details of the instrumental hardware. When the data are correlated, the expected delay is calculated, and the data streams from each element are aligned at that delay. For spectroscopy, a correlation function centered on the expected delay is formed and Fourier transformed, to obtain the cross-power spectrum. For continuum work, if the delay is known *a priori* to much better than the inverse of the bandwidth, only the correlation at the expected delay need be determined. The delays are usually not known well *a priori*, so the correlation function must be determined over a wide range of delays to ensure that the signal is retained. In many cases, it is useful to transform the correlation function into a frequency spectrum, in which the delay error appears as a phase slope. The spectrum takes less space, as only one sideband is kept, and is what is needed for global fringe fitting.

Significant delay errors can lead to large phase offsets, since the phase is the delay times the observing frequency (which is large). Phase offsets, as such, are not a problem because they will be corrected by self-calibration. But if they change rapidly compared to the desired integration time, the data will be degraded. The rate of change of the difference between the expected and actual phases is known as the *residual fringe rate*. The integration time must be kept well under the inverse of the residual fringe rate.

For VLBI, the use of separate clocks, uncertainties in the geometry and atmosphere, and high resolution (i.e., high sensitivity to uncertainties in geometry) combine to make accurate *a priori* determinations of the delay and phase difficult. The troposphere and ionosphere, the relative drift of the clocks, and even the position of the Earth all change on day-to-day (or shorter) time scales in ways that are hard to predict. The correlator must therefore deliver several delay (or frequency) channels, and records integrated for only a few seconds. To keep the volume of data reasonable, either special observations must be made to determine clock and geometric parameters (consuming much of the observing time) or the delay and fringe rate must be determined from the data by fringe fitting—the traditional method. Note that fringe fitting (like self-calibration) removes absolute position information, because the geometric information is contained in the delays. The delays determined by sophisticated versions of fringe fitting are, however, used for VLBI astrometry and geodesy.

4.1. Theory of fringe fitting.

Conceptually, fringe fitting is simple. The correlator delivers a correlation function as a function of time, covering a delay range larger than any delay uncertainty. These data can be thought of as a two-dimensional matrix with delay and time axes. The trick is to find the signal. If the time axis is transformed into a frequency axis (*residual fringe frequency*, often called *fringe rate*), the signal will appear as an isolated peak in the matrix. The simplest versions of fringe fitting merely search for, and fit for, the parameters of the peak in the data from each baseline, for each time interval of some specified length. Examples of

Figure 19–1. Plots of residual fringe frequency spectra for 7 delay lags for **(a)** a strong source and **(b)** a weak source, after an unusually long, one-hour integration. Note how the signal appears in adjacent lags at the same frequency. For weak sources, the problem is to distinguish the signal from noise spikes. In these plots, the fringe frequency window is 50 mHz wide in (a) and 25 mHz wide in (b). This was an early VLBI detection of SS 433.

residual fringe frequency spectra for several delay lags are shown for a strong source in Figure 19–1a and for a weak source in Figure 19–1b.

Finding the signal is easy for strong sources, but for weak sources—often where the interesting science lies—it can be tricky. Fringe fitting methods are being actively developed, and the following describes the situation at mid-1988.

Most of the uncertainty in delay and phase, except for the part due to source structure, stems from uncertainties in parameters for each interferometer element. The measured visibilities, in frequency-time coordinates, can be written as

$$\widetilde{V}_{ij}(t,\nu) = g_i(t,\nu) g_j^*(t,\nu) V_{ij}(t,\nu) + \epsilon_{ij} \,, \tag{19–1}$$

where $V_{ij}(t,\nu)$ is the true visibility of the source on the i–j baseline at time t and frequency ν, $g_i(t,\nu)$ and $g_j(t,\nu)$ are the antenna voltage gains,

$$g_i(t,\nu) = a_i(t,\nu) e^{i\phi(t,\nu)} \,, \tag{19–2}$$

and ϵ_{ij} is a thermal noise term (see also Lecture 9). The clock, geometric, tropospheric, etc., uncertainties are absorbed into phase slopes in the gains. Equation 19–1 explicitly assumes closure by asserting that the baseline gain can be expressed as the product of antenna gains. A non-closing term should be included for each baseline, but in practice such terms are less than a few percent and can be ignored except in high dynamic range imaging. In the following equations, the noise term is also ignored—just remember that any measured quantity includes a contribution from noise.

Fringe fitting involves only the phases of the visibilities. Amplitudes are calibrated by more traditional means, as discussed later. Separating the amplitude and phase parts of Equation 19–1 and dropping the (t,ν) dependence gives

$$\widetilde{A}_{ij} e^{i\widetilde{\theta}_{ij}} = a_i a_j A_{ij} e^{i(\phi_i - \phi_j + \theta_{ij})} \,, \tag{19–3}$$

or for the phases alone:

$$\widetilde{\theta}_{ij} = \phi_i - \phi_j + \theta_{ij} \,, \tag{19–4}$$

where $\widetilde{\theta}_{ij}$ and θ_{ij} are the measured and true visibility phases on the i–j baseline—both are functions of time and frequency. The reason for the term "closure" can be seen by summing the phases around a closed triangle of antennas i, j, k. All the ϕ_i, ϕ_j, ϕ_k's appear twice with opposite sign, and cancel, so we have

$$\widetilde{\theta}_{ij} + \widetilde{\theta}_{jk} + \widetilde{\theta}_{ki} = \theta_{ij} + \theta_{jk} + \theta_{ki} \,. \tag{19–5}$$

At this point, a simplifying assumption will be made—that, for the bandwidth and time interval of the fit, the amplitudes of the gains are constant, while the phases vary linearly with both time and frequency:

$$\begin{aligned} \widetilde{A}_{ij} e^{i\widetilde{\theta}_{ij}} &= a_{i0} a_{j0} A_{ij} \times \exp\, i\Big[\phi_i(t_0,\nu_0) - \phi_j(t_0,\nu_0) + \theta_{ij}(t,\nu) \\ &+ \big(r_i(t_0,\nu_0) - r_j(t_0,\nu_0)\big)(t - t_0) + \big(\tau_i(t_0,\nu_0) - \tau_j(t_0,\nu_0)\big)(\nu - \nu_0)\Big] \,, \end{aligned} \tag{19–6}$$

where the r_i are the antenna fringe rates,

$$r_i = \frac{d\phi_i(t_0, \nu_0)}{dt}, \tag{19–7}$$

and the τ_i are the residual antenna delays expressed as phase slopes,

$$\tau_i = \frac{d\phi_i(t_0, \nu_0)}{d\nu}. \tag{19–8}$$

The purpose of fringe fitting is primarily to determine the τ_i and r_i, so that the delay and rate offsets can be removed, and the data can be averaged in both time and frequency. For a single baseline, the net delay and fringe rate can be determined by fitting phase slopes in frequency and time or, as mentioned earlier, by transforming to delay-rate space and looking for a peak.

Since all the gain terms are antenna-dependent, it is possible, as in self-calibration (Lecture 9) to use all the data to determine the antenna gains. This allows greater sensitivity, lets us use data from stronger baselines to determine the parameters for weaker baselines, and ensures that the delays and rates used really do close and do not introduce unnecessary noise in the imaging results. The process of solving for antenna-dependent delays and rates is called *global fringe fitting.*[4] It is described in detail by Schwab and Cotton (1983). As in any self-calibration scheme dealing with phases, there is enough information to establish the parameters of $N-1$ antennas relative to a reference antenna, where N is the number of antennas in the array. The phase, rate, and delay of the reference antenna must be set arbitrarily—usually to zero.

Global fringe fitting may be done either by a least-squares solution or by a Fourier transform solution. In practice, the Fourier transform solution is often used to give an initial guess for the least-squares solution. In the least-squares solution, the antenna phases, fringe rates, and delays are determined using data from all baselines assuming a source model. Equation 19–6 is used, with the visibilities for the assumed source model substituted for the true visibilities. For details, see Schwab and Cotton (1983). The quality of the fit depends on the signal-to-noise ratio and on the quality of the model. If the model is bad (radians of phase error), it may be hard to find a fit and the number of baselines may have to be restricted, as discussed below.

For the Fourier transform solution, the frequency-time data are transformed to the delay versus fringe rate domain. The fringes will appear at an isolated point in this domain, so, to locate the peak, we need only find the maximum and fit adjacent points. The complication comes in setting up a frequency-time matrix that uses data from more than just the baseline between the antenna being calibrated and the reference antenna. Useful information can be obtained from the observed phase on the direct baseline and from the sum of the phases from any group of baselines that connect the two antennas via other antennas. Only the one, two, and three baseline combinations are independent. Combinations

[4] Global fringe fitting is done in AIPS by the program 'CALIB'.

of larger numbers of baselines can be expressed as sums of smaller combinations. Even the two and three baseline combinations are not fully independent because individual baselines contribute to several such combinations, but each contains some information not available in any other. The equations for the phase differences are, in the three cases:

$$\begin{aligned}
\phi_{ij} &= \phi_i - \phi_j = \widetilde{\theta}_{ij} - \theta_{ij}\,, \\
\phi_{ikj} &= \phi_i - \phi_j = (\widetilde{\theta}_{ik} - \theta_{ik} + \phi_k) - (\theta_{kj} - \widetilde{\theta}_{kj} + \phi_k) \\
&\qquad\qquad = (\widetilde{\theta}_{ik} + \widetilde{\theta}_{kj}) - (\theta_{ik} + \theta_{kj})\,, \\
\phi_{iklj} &= \phi_i - \phi_j = (\widetilde{\theta}_{ik} - \theta_{ik} + \phi_k) - (-\widetilde{\theta}_{kl} + \theta_{kl} + \phi_k - \phi_l) - (\theta_{lj} - \widetilde{\theta}_{lj} + \phi_l) \\
&\qquad\qquad = (\widetilde{\theta}_{ik} + \widetilde{\theta}_{kl} + \widetilde{\theta}_{lj}) - (\theta_{ik} + \theta_{kl} + \theta_{lj})\,,
\end{aligned} \tag{19–9}$$

where the j^{th} antenna will usually be the reference. The phases of the intermediate antennas along any string of baselines enter the sum twice, but with opposite signs, so they cancel. Each combination depends only on measured and true visibility phases.

A frequency-time data array can be formed with the complex, weighted sum,

$$F_{ij} = w_{ij} e^{i\phi_{ij}} + \sum_k w_{ikj} e^{i\phi_{ikj}} + \sum_{k,l} w_{iklj} e^{i\phi_{iklj}}\,, \tag{19–10}$$

where the weights reflect factors that may include the signal-to-noise ratios of the baselines, preconceived notions of the value of certain stations or baselines, and the expected source visibility amplitude on the baselines. Using a reduced set of baselines is equivalent to setting some weights to zero. Again, for details see Schwab and Cotton (1983). F_{ij} is a frequency-time matrix that can be Fourier transformed in both dimensions and searched for a peak in the resulting delay-fringe rate matrix.

Note that to use Equations 19–9 and 19–10, both the measured visibility phases and estimates of the true visibility phases are needed. Without the latter, a complicated source could introduce such large phase differences that the terms in Equation 19–10 would cancel each other rather than enhance the signal-to-noise ratio. If the source structure is unknown, as it usually is on the first pass (why is it being imaged?), a restricted set of baselines may be used to solve for each antenna's parameters. For example, we may use just the direct baseline to the reference antenna, or the two baseline combination through some other strong station. Errors in the assumed true visibility phases that do not degrade the amplitude of the sum (i.e., are less than about a radian) will not seriously degrade the solution for fringe rate and delay, however. As the visibility phases due to structure usually vary slowly in time and frequency, the errors in the model introduce errors mainly in the phase, but not in the fringe rate and delay. Phase errors are removed in the later imaging stages. For these reasons, the initial model need not be especially good. These considerations also apply to the least-squares solution.

The result of fringe fitting is a table of the delay, phase, and fringe rate offsets for each antenna relative to the reference antenna, as functions of time. The table can be used to correct the measured data. Once the residual delays and fringe rates (and phases too, although this is not especially important) are removed, the data can be averaged. The limits to the delay averaging are set by the field of view or the desired spectral resolution. The limits to time averaging are set by the field of view (Lecture 13), the coherence time, and/or the scan lengths. The averaging time should probably be kept shorter than, or equal to the time interval used in the fringe fit, as discussed below.

Typical VLBI network experiments use antennas with widely differing sensitivities. If a source can be detected at all on most baselines, the signal-to-noise ratio must then be high on a few baselines. Alef and Porcas (1986) developed a fringe fitting scheme that exploits this, while reducing the computer resources needed for global fringe fitting. They make traditional, single-baseline fringe fits to all the data, or to a subset containing at least one baseline per station, storing only one baseline's data at a time on the computer. Once all the fits are made, the results from all baselines are combined to obtain the station-dependent delays and rates used to correct the data. Except when a weak source is observed with similar antennas (probably a common case for the VLBA), the results should be similar to those obtained by global fringe fitting.

4.2. Practical considerations.

So much for the theory of fringe fitting—now for the art. For several reasons, fringe fitting is not straightforward. We must decide which baselines to use in a solution, how to weight the baselines (or antennas), what model to use, how long an interval of data to use in the solution, what reference antenna to use, etc. The answers are not always clear. Many of these questions arise because of the wide range in sensitivity of the baselines in typical VLBI network experiments, resulting from the use of a wide variety of antennas. Resolution effects also increase the range of signal-to-noise ratios on the different baselines, often in a time dependent manner. In many VLBI experiments, no one antenna is on for the entire observation. At least two reference antennas must then be used, and glitches at the time of transition must be avoided.

If the program source has unknown structure that is likely to cause the true visibility phase to be more than about a radian from zero it is probably best to use only single baselines to solve for the antenna delays and rates. If the signal-to-noise ratio on the single baselines is low, the multiple baseline combinations can be used, while accepting that some data will be lost. An image can be made from the results of this preliminary fit and used in a second fit, presumably with better results. If the source is dominated by a point component, so that the true visibility phases are only a few tens of degrees on any baseline, any baseline combination can be used.

Second fits should be avoided, if possible, because fringe fitting demands large computer resources. The object is to determine the delay and fringe rate accurately enough that the signal is not degraded by being averaged in delay and time. At high signal-to-noise ratio, a single pass with single baseline solutions and a point source model may suffice. No matter how bad the model is, if good signal-to-noise ratio is obtained in the fit, the only significant errors will be in the phases—and those can be fixed later by self-calibration. A second fit is not needed.

How well should the delay be determined? The amplitude, as a function of delay, will roughly follow a sinc x function (roughly—because the bandpass is not perfectly square). For Nyquist sampling, the spacing of delay channels is half the inverse of the bandwidth. The first null in the ideal sinc x occurs two lags from the true delay. The amplitude will then be degraded by 1%, with a delay error of about 1/6 of a lag. Delay errors can cause closure errors, so large errors should be avoided if possible. For most experiments, delays good to 1/6 of a lag (about 40 nsec for Mk II) are a reasonable goal. Note that the tolerable delay error depends on the distribution of base-band channels in multiple band systems like Mk III and must be calculated based on the experimental setup.

The optimal way to use the delays derived in fringe fitting is currently under study. The fringe fitting program delivers a table of delays and fringe rates that, in the default mode, are interpolated and used to correct the data. The results are noisy, however, especially on weak sources and/or weak antennas. The causes of the delay errors, such as clock offsets and geometric and tropospheric uncertainties, are likely to be smoothly varying. The scatter in the measurements is mainly due to measurement noise, not to real fluctuations in the delay errors. It is therefore reasonable to smooth the fits before applying them to the data, but care must be exercised if delay steps were used in the processing, or if there are times when the delay changed rapidly (e.g., because of rapid and unmodeled tropospheric variations in low elevation data).

For a recent 18-station experiment aimed at very high dynamic range, the fringe fit results were hand-smoothed, to test the procedure. The results are encouraging, so automatic fit smoothing should probably be provided. Figure 19–2 shows the residual delays and fringe frequencies, as output by the fringe fitting program, for some of the stations of this 18-station experiment. The differences in signal-to-noise ratio between various stations are clear, as are the jumps at times when the reference antenna had to be changed because the source was not above the horizon at any one antenna for the entire observation (note that the time ranges of the various plots are not aligned). The delays were smoothed by removing any steps introduced by clock offset changes used during correlation to center the fringes in the delay window (some clocks drift rapidly); by applying an offset at the reference antenna change to remove that step; by generating a smooth delay for each station consisting of a few straight lines approximating the fit results; by re-introducing all the steps; and by using the final results to correct the data. The fringe frequency fits were not smoothed. Examination of the fit results using plots such as those in Figure 19–2 is a good way to assess the quality of the fit and to determine what corrections should be made. Software to display and manipulate such fits is still being developed.

With the wide range of sensitivities on the current VLBI networks, it is often best to fit only the baselines to a sensitive antenna (e.g., VLA or Bonn), that is used as the reference. The high signal-to-noise ratio on such baselines means

Figure 19–2. Plots of the fitted residual delays **(a)** and fringe frequencies **(b)** for 6 antennas of an 18-antenna experiment. Note that the time scales are not aligned. The reference antenna changed at about 1.75 days, causing a jump in delay and rate. Other large delay jumps were caused by steps in the assumed clock offset used during processing to keep the fringes centered in the correlator delay window. The large changes in fringe frequency at extreme times are caused by unmodeled atmospheric effects at high zenith angles.

that the others contribute little to a many-baseline fit. Problems may arise if the source is heavily resolved on the long baselines; we may then have one good reference antenna for U.S. baselines, and another for European baselines, but none for the whole array. We might then use two-baseline combinations, rather than the direct baselines. For weak antennas far from the reference antenna, the high signal-to-noise ratio baseline to the nearby sensitive antenna could be used, along with the baseline between the sensitive antennas, to derive the fit results.[5]

Another choice that must be made for fringe fitting is the time interval over which to do the fit. The longer the time interval, the higher the signal-to-noise ratio. The linearity assumption of Equation 19–6 eventually fails however, either because some offsets due to the troposphere change rapidly or because the frequency standards fluctuate. We refer to this situation, generally, as *loss of coherence*; its effect on an integrated signal is to reduce the amplitude and spread the power in fringe rate. For imaging, good amplitudes are important, so the integration time after fringe fitting should be much less than the coherence time. Amplitudes are not important for fringe fitting, however, so integration can be extended for as long as the signal-to-noise ratio increases. As the noise decreases as the square root of integration time, but the loss of coherence reduces the amplitudes more slowly at first, it is often worth integrating longer for fringe fitting than for imaging.

The choice of integration time therefore depends on the signal-to-noise ratio, on the scan lengths, and on the maximum set by coherence. If the signal-to-noise ratio is high, short integrations are preferable, to avoid averaging over tropospheric fluctuations. The fit interval should also be a reasonable match to the length of the observing scans. Values that give short effective fit intervals at the ends of scans should be avoided, as the low signal-to-noise ratio in these intervals degrades the results. In present VLBI networks, the maximum integration time allowed by coherence is often set by one or two antennas that have especially poor frequency standards. In practice, the fit interval can be selected by fringe fitting a short subset of the data, using several different intervals, and choosing the one that works best. This is also a good way to decide which reference antenna and baseline combinations to use. Getting a "feel" for the data in this way may avoid problems, and possible reprocessing, later on.

5. Amplitude Calibration

For calibration, we wish to use Equation 19–1, along with knowledge of the antenna gains, to estimate the true visibility amplitude from the data. Rewriting Equation 19–1, retaining only the amplitude terms of the gains (Eq. 19–2) and ignoring the noise, we have

$$S_{ij}(t,\nu) = \frac{\widetilde{A}_{ij}(t,\nu)}{a_i(t,\nu)a_j(t,\nu)}\,, \tag{19–11}$$

[5] The AIPS program 'CALIB' tries to find fringes to a secondary reference antenna if the baseline to the reference antenna has fringes with too little signal-to-noise during single baseline fits.

where S_{ij} is the calibrated, correlated flux density and $\widetilde{A}_{ij}$ is the raw correlation coefficient from the correlator. The correction factors that account for the signal-to-noise ratio losses due to effects such as digitization have been absorbed into the a_i. It is instructive to rewrite Equation 19–11 in a form that applies to one frequency and time and that shows how the various calibration parameters are used:

$$S_{ij} = \widetilde{A}_{ij} b \sqrt{\frac{T_{\mathrm{sys}_i} T_{\mathrm{sys}_j}}{K_i K_j}}\,, \qquad (19\text{–}12)$$

where b is the factor that accounts for digitization losses etc., the K_i are the antenna sensitivities in K Jy^{-1} (Lecture 7), and the T_{sys_i} are the system temperatures in Kelvins.

Most current experiments use Equation 19–12 directly. The value of b depends on the encoding system and on details of the correlator. It is typically about 2.5. The system temperatures are measured during the observations, and gain curves are determined at some other time. One complication of current VLBI is that stations provide the necessary information in different forms, and with different levels of reliability. For strong sources, for which antenna temperatures can be measured, the K_i can be replaced with $T_{\mathrm{ant}_i}/S_{\mathrm{tot}}$, where S_{tot} is the source total flux density and T_{ant_i} is the antenna temperature. While there is an effort to provide system temperatures continuously, most observatories still measure them only sporadically. When there are only a few measurements, it is advisable to screen them for bad values.

The measured system temperatures and gain curves rarely provide calibration consistent to better than about 10 percent. To improve the calibration, one or more calibration sources expected to have simple structure and high signal-to-noise ratio are usually observed. Models and/or images of these sources are used with self-calibration to determine constant offsets for each station and thus to improve the *a priori* calibration of the program source data.

6. Continuum Imaging

After fringe fitting and amplitude calibration, the data set is similar to one from a connected-element interferometer that has not been phase calibrated. Image making begins with selecting a source model to use in the first pass of self-calibration (Lecture 9). A point source model often works, and some observers never use anything else. A more complicated model may help when the data are limited (but this is controversial). Some self-calibration software apparently works better than others when given a poor starting model.[6] If a more complicated model is needed, it can be obtained by fitting several Gaussian components to some or all of the visibility data, but model fitting itself requires an initial guess. This guess can come from examining plots of the visibility data and estimating what structure would produce the observed minima and maxima.[7] Deep minima in the visibilities aid this process by showing the position angle

[6] NRAO's AIPS routines seem to work with a point source starting model.

[7] Note that model fitting has uses other than to provide a first guess for self-calibration. When the source structure is simple and unconfused, it is easier to estimate errors by fitting source

and separation of features, but the problem is difficult if the source structure is complicated. No recipe is known that always works reliably. Because there is no preliminary phase calibration, and the data sets are often small (providing few constraints), this step is much harder for VLBI than for interferometry with connected elements.

Imaging VLBI data often involves many passes through the self-calibration and re-imaging cycle. At each pass, the previous image is used as a better starting model for the self-calibration. In effect, the antenna gains and the source structure are iteratively solved for simultaneously. Because the procedure is iterative, there is no need to get everything right on the first pass, and attempts to do so are often counterproductive. The amplitudes should not be self-calibrated for the first several iterations; the *a priori* amplitude calibration is probably better than what self-calibration provides if the model is poor. It is best to wait until the phase calibration is converging well before tackling the amplitudes. Also, as discussed in Lecture 16, it is usually worth restricting the range of u-v data used in the self-calibration. This is obviously worthwhile when a large, complex source is dominated by a point component. The initial fit should use only the longest baselines, which respond primarily to the point component. As the image improves, data from shorter baselines can be used and the self-calibration can be extended to the amplitudes—first, to estimate constant scaling factors for each antenna, and later to determine time variations. The full range of u-v spacings may never be used, especially if the shortest baselines have excess correlated flux density from poorly modeled, extended structure.

A point source starting model biases the self-calibrated image toward symmetrical sources. Artifacts that may result from this are false "counterjets" in core-jet sources and triple structure in double sources. This bias may be minimized by restricting the fits to long baselines at first, but the full procedures needed to avoid such behavior are not yet clear. As the number of baselines is increased, the number of independent constraints increases rapidly, so this problem becomes less severe.

It is essential to be alert to possible false features or other problems with the images. There are two primary clues to the credibility of the image: its appearance, and the ability of the significant features of the image (e.g., those contained in the 'CLEAN' components) to predict the data. The appearance can be judged by looking at contour plots or TV images. If there are regular artifacts (ridges, etc.), problems likely remain. Negative holes with amplitudes significantly above the noise are a good indicator of trouble (unless, of course, they are expected, as in spectral absorption experiments). If there are any dubious or controversial features, a serious effort should be made to make them go away. If a feature persists no matter what is done, it is more likely to be real. The apparent noise level may also indicate quality, especially in high dynamic

---

parameters in the visibility domain than by fitting them in the image plane, especially when estimating the deconvolved size of a barely resolved feature. A feature that resembles the beam in the image plane may have obviously reduced visibility amplitudes in the u-v data. However, any extraneous structure also affects the u-v data—if such structure is present, it is best to extract parameters by fitting models to the final image.

range images. If successive iterations do not reduce the noise level, the process should be stopped or perturbed. But the noise level alone is not a reliable indicator of quality—apparently low-noise images may contain spurious features and not reproduce the u-v data well.

The ability of the model to reproduce the u-v data has been a traditional quality indicator for VLBI—the VLBI literature contains endless plots of models overlayed on data. With small quantities of data, it is possible to obtain images that seem reasonable but whose associated models do not reproduce the data well because, at low dynamic range, significant signal may be generated by spurious features of the residual image. With large data sets and high dynamic range, the data should be reproduced better. When dynamic ranges exceed 100 : 1, improvements in the images may predict only minor changes in the u-v data. These changes can be hard to detect in u-v plots, but the fit of the model to the data can reveal problems (see Fig. 19–4).

A factor that can be adjusted during image deconvolution with the 'CLEAN' algorithm is the size and placement of the 'CLEAN' window(s) (see Lecture 8). If there is an *a priori* reason, such as previous images of the same source, to believe that a source is confined to a certain region of sky, imaging can be guided by confining the 'CLEAN' windows to that region. This is especially important with small, poor-quality data sets. If the 'CLEAN' window is too small, it is hard to get a good image. 'CLEAN' will try to account for the flux density outside the window by putting negative features and other obvious artifacts in the window. I know of an example where an observer could not get a good image until he saw another image of the same source that had extended emission well outside his 'CLEAN' window. When he opened the window in the correct direction, a good image was obtained quickly.

Two other parameters that can be varied efficaciously during VLBI imaging are the reliability weights R_k and the density weighting function D_k (see Lecture 6). If all visibilities are weighted equally, and uniform density weighting is used, sidelobe levels will be low but the noise will be high (because of the high noise on the least sensitive baselines). This is not a problem for an instrument with similar antennas, but for present VLBI experiments the effect is severe. The lowest noise is obtained if the data are given reliability weights that reflect their signal-to-noise and natural density weighting is used, but this produces high sidelobes if there are only a few sensitive baselines. The best approach may be to image with level reliability weights and uniform density weighting while the errors are dominated by phase calibration errors (the first iterations of self-calibration or hybrid mapping). Then, as the image improves and the noise on the weak baselines becomes significant, one may move gradually toward signal-to-noise reliability weighting and natural density weighting. If the calibration is good, the deconvolution procedures should handle the high sidelobes—which they cannot do while the calibration is poor. This area needs further study.

Several antenna-dependent factors can also be used to control self-calibration. One is the weight given to each antenna, based on, for example, the reliability of the *a priori* calibration (mostly useful for amplitudes). The range over which the amplitude calibration of each antenna is allowed to vary can also be restricted, so that some antenna gains are constrained to remain nearly at their *a priori* values while others vary. Given the wide range in the quality of the *a priori* calibration now typical in VLBI, this can be very useful. Finally, the smoothing time-scale for the amplitudes may range from "point to point" for some antennas while others are allowed to change only slowly, again using information on the quality of the calibration of each antenna.

Clearly, many parameters can be adjusted while attempting to obtain a good image. No single procedure is guaranteed to work, and each new case may require a slightly different method. Experience helps, but sometimes false starts and many iterations are made before getting a good image—while other times convergence is rapid, with minimal fiddling of parameters. More advanced algorithms that make VLBI imaging easier and more reliable are sorely needed. New algorithms should be tested by imaging fake data sets (for which the true image is known, but not by the person doing the imaging) to obtain a feel for how well they work in practice.

Figure 19–3 illustrates several aspects of the VLBI imaging process. Figure 19–3a is the result of the first pass of self-calibration with a point-source starting model. The image is obviously poor, but a definite position angle has been established and some asymmetry has been introduced. Figure 19–3b shows the result after 11 iterations. The source is now much more compact and asymmetric. The noise level is lower (note the lowest contour level), but there are still hints of problems—uncomfortably deep negative features to the east and suspicious, low-level features off the line of most of the others. Comparison with images made at other times also casts suspicion on the eastern-most feature that appears as a bulge on the side of the strongest feature.

As the appearance of the image in Figure 19–3b is good relative to many VLBI images, there is a temptation to stop and adopt it as the final result. As it turns out, that would be a mistake. The eastern-most feature (which might be treated as the core and so be used to align this image with those from other epochs) is not real. The best indication of a problem at this stage was not in the image itself, but in a comparison between the predicted and calibrated data. Figure 19–4 shows the calibrated phases (crosses) and predictions of the 'CLEAN' components of the image (smooth lines). On one baseline (Haystack to Owens Valley, denoted HSTK–OVRO), the prediction misses the data by a small but *significant* amount. Two other baselines, not displayed here, showed smaller problems. These offsets could not be removed by further self-calibration. Closer examination of the fringe fitting revealed that the clock at one station (not HSTK or OVRO!) was poor, and too long an integration time had been used. There were further problems when the raw fringe rate or the delay rate was near zero. After re-fitting the fringes and further editing, the much improved image of Figure 19–3c was obtained. It shows weak features far from the main emission region that are confirmed by observations at other frequencies. It also contributes to a consistent picture of the evolution and motions of components in the source. The improvement in the fit of the predicted data to the calibrated data can be seen clearly in Figure 19–5.

The data shown in Figure 19–4 illustrate the problem of the wide varia-

Figure 19–3. VLBI images of a compact continuum source (3C 120) at three stages of the imaging process. **(a)** The first iteration. One phase-only self-calibration has been made using a point source model. The contours are at −35, 35, 69, 139, 208, 278, 347, 417, 486, 556, and 625 mJy/beam. **(b)** After 11 iterations. This was as good an image as could be made without fixing some problems in the data as discussed in the text. The extension to the east (left) of the brightest feature is not real. The contours are −10, −5, 5, 10, 20, 30, 40, 50, 60, 70, 80, 100, 150, 200, 250, 300, 350, 450, and 550 mJy/beam. Note that the lowest contour in (a) is between the fourth and fifth contour here—significant progress has been made. **(c)** The final image made after fixing problems in the data and doing many more iterations of the self-calibration loop. The image is now much better, and significant structure has appeared to the west of the bright regions. The fine details of this structure are in some doubt, but general features such as the emission feature at about 0.05 arcseconds are confirmed by independent observations at a lower frequency. The contour levels are −4, 4, 8, 13, 19, 27, 37, 52, 72, 100, 139, 193, 268, 373, 518, and 720 mJy/beam.

Figure 19–4. A display of data phases and phases predicted by the 'CLEAN' components of the image of Figure 19–3b for 6 baselines. The crosses represent the data, and the smooth curves are the model. Note the systematic offset on the HSTK–OVRO baseline. Also note the wide range of signal-to-noise ratio on the various baselines.

tion in signal-to-noise ratio for VLBI data. Clearly the HSTK–VLA data were much better than the NRL–HRAS (Maryland Point to Fort Davis) data at this frequency. When phases scatter as much as they did here on the NRL–HRAS baseline, it may be best to edit out the baseline entirely.

Figures 19–3 and 19–4 show an extreme example of trouble encountered during imaging despite having abundant data—Mk II observations of a 3 Jy source from 11 stations. It should have been easy. Clearly the procedures are not yet routine.

7. Other Hazards

The digital fringe rotation and delay setting degrade the signal-to-noise ratio, as discussed in Lecture 4 and above. Unless corrections are made, they can cause much worse problems on short baselines and at times when either the raw fringe rate or delay update rate are small. I refer here to the total rates

Figure 19–5. A display of data phases and the phases predicted by the 'CLEAN' components of the image shown in Figure 19–3c. Note the considerably improved fit over the data shown in Figure 19–4. The NRL–GRAS baseline was deleted for lack of usable signal. Even at this stage there are a few bad points that could be edited out for a small improvement. However, a few bad points in a large data set usually have little effect on the image. This plot appears different from that in Figure 19–4 because it was produced with the Caltech VLBI package rather than with AIPS.

removed by the processor, not the residual rates discussed under fringe fitting. If the fringes go through less than a few turns, or the delay change is less than a few bits during an integration time, the degradations differ from the statistical effect at large rates. The magnitude of the offsets resembles a $\operatorname{sinc} x$ law, so it takes several turns or updates before they are insignificant. Consider zero fringe rates or constant delays—there is no degradation, so when such data are combined with data from long baselines, there can be large closure errors. The effects are calculable, so data in these dangerous regimes should be corrected or deleted. For the VLBA, all data will be corrected. Baselines shorter than about 500 km may have problems for significant lengths of time in 18 cm Mk II observations, and the effects scale with frequency and bandwidth. Higher frequency observations have higher fringe rates, so slow fringe rate problems occur only on shorter baselines. Wider bandwidth observations have faster delay update rates, so the slow delay-rate problem occurs only on shorter baselines. Note that the crucial bandwidth is that determining the sampling rate—for multi-band data such as Mk III, the effect depends on the individual channel bandwidth.

8. Polarization Studies

To determine the polarized brightness distributions of a radio source in more than one of the four Stokes parameters, each antenna should be outfitted with two receivers[8] and feed systems whose responses should be, as nearly as possible, mutually orthogonal.

When linearly polarized feeds are used with widely separated interferometer elements (implying large parallactic angle differences), both the cross hand correlators and the parallel hand correlators respond significantly to each Stokes parameter (*cf.* Thompson, Moran, and Swenson 1986). However, when all antennas are equipped with dual circular polarization feeds and the fractional linear polarization is small, the parallel hand correlators respond primarily to Stokes I and V.[9] Thus, VLBI software engineering is easier with circularly polarized feeds.[10] There are no major theoretical or practical obstacles to dealing with the other case, except that current software packages used in VLBI (notably AIPS) have been designed for the circular–circular case which requires, e.g., in self-calibration and the global fringe search for the total intensity data, only the parallel-hand data.

A problem of polarimetry with the present VLBI networks is that not all stations can observe in dual polarization mode, and, among those that can, some have linear-response feeds and others have circular-response feeds. The (im)purity of the feed responses always is such that useful results cannot be obtained without painstakingly careful calibration. One of the best-equipped and most sensitive telescopes (the 140-ft in Green Bank) is equatorially mounted, and the largest (Arecibo) has many peculiarities; simple gain calibration, let alone polarization calibration, thus is a problem. The data reduction must account for differences in the parallactic angles between widely-separated antennas, and possibly for variation of the parallactic angle at alt-azimuth mounted antennas while the angle is constant at equatorially mounted ones. Most widely available software is inadequate for polarization calibration.[11] Polarization calibration of VLBI data is discussed by Roberts and Wardle (1986). Present VLBI polarization studies are therefore carried out with arrays consisting, typically, of only

---

[8] This is not an absolute requirement but is highly desirable. If not all antennas can observe in dual polarization, then the point-source response functions for one's images are complex, and a complex deconvolution (deconvolution of one complex function from another) is required. The same is true if (as frequently happens in VLBI) data are sporadically missing from one of the two receivers on an antenna. Although there is no fundamental theoretical or computational obstacle here, standard software packages, such as AIPS, do not include a complex deconvolution capability. Cotton *et al.* (1984) and Roberts *et al.* (1984) present VLBI polarization results obtained via complex deconvolution.

[9] The parallel-hand correlators respond only to I if $V \equiv 0$ and if the feeds are perfectly (and identically) circularly handed, so either parallel-hand correlation can be used to estimate the total intensity in this case.

[10] All feeds for the VLBA are designed to respond to pure left- or right- circular polarization states, so far as is economically achievable.

[11] AIPS handles only the circular–circular case; private software is used for linear–linear; and mixed linear–circular data are avoided.

four to six elements. The global fringe search algorithm helps polarization studies considerably because it disposes of the requirement (of baseline-by-baseline fringe searching, with circularly polarized feeds) that the source be detectable in the cross-hand channels—i.e., the total intensity information can be used to obtain the residual delay and residual delay rate (fringe frequency) information relevant to the cross-hand channels.

Radio polarization data may be severely corrupted by ionospheric Faraday rotation, especially near dawn and dusk. This effect is seen nearly always during daylight at wavelengths longer than 18 cm, and often, during solar active periods, at wavelengths as short as 6 cm. It can sometimes be corrected—e.g., by satellite-aided monitoring of the ionosphere or by monitoring of polarized calibrators. Often, however, polarization data must simply be discarded or ignored. Because of the λ^2 dependence of the effect, *all* radio astronomy polarization measurements are much easier to carry out at wavelengths shorter than 18 cm. Its consequences for VLBI, where the effects will be different at a given instant for different antennas, have yet to be investigated thoroughly. The extent to which this effect can be corrected in VLBI, via self-calibration and careful monitoring of the ionospheric total electron content at one VLBI station, is unclear.

For these reasons, even though VLBI is now in its third decade, polarization studies are done mainly by a few specialist teams, and little of their work has yet been published. Cotton *et al.*'s (1984) observations of 3C 454.3 were the first published VLBI linear polarization images. Roberts, Gabuzda, and Wardle's (1987) observations of the BL Lac object OJ 287 revealed complex, time-variable structure in the linear polarization (on the milli-arcsecond scale) that is not at all evident in total intensity maps, and in their polarization data they find evidence of superluminal motion.

Many of the current difficulties in VLBI polarization studies (which are exacerbated by limitations of the analysis software) arise from inhomogeneities in the instrumentation. The VLBA, we hope, will cure this ill.

9. Conclusion

This lecture has emphasized the differences between VLBI and connected-element interferometry. Most of the differences involve added complications for VLBI, and added ways in which the data reduction can go awry. But I do not want to leave the impression that VLBI images can be made only by experts and cannot be trusted. The techniques of VLBI are not fundamentally different from those of other interferometry, so any knowledgeable and careful VLA user should be able to do VLBI. I emphasize again the need to assess the reliability of the final images, remembering that the appearance of an image alone may not be a reliable quality indicator, especially with small amounts of data. It is usually worth checking the fit of the significant parts of the image (e.g., 'CLEAN' components) to the data. If anything looks suspicious, vary the imaging parameters, or start over with a different initial model or u-v range, to test how well features repeat.

That carefully constructed VLBI images are reliable is shown by many cases where images, made at different times with different arrays, show either the same structure or a simple evolution of the structure. The reward of VLBI is the high resolution that can be obtained—higher than with any other imaging technique at any wavelength. That resolution allows one to look deep into the energetic regions of many objects and to watch the time evolution of structures in even the most distant quasars on human time scales. As we move toward the era of the VLBA, we hope that many of the difficulties of doing VLBI will be eased and that the community of astronomers who make VLBI observations at least occasionally will expand greatly.

References

Alef, W. and Porcas, R. W. (1986), "VLBI fringe-fitting with antenna-based residuals", *Astron. Astrophys.*, **168**, 365–368.

Cotton, W. D. (1979), "A method of mapping compact structure in radio sources using VLBI observations", *Astron. J.*, **84**, 1122–1128.

Cotton, W. D., Geldzahler, B. J., Marcaide, J. M., Shapiro, I. I., Sanromá, M., and Rius, A. (1984), "VLBI observations of the polarized radio emission from the quasar 3C454.3", *Astrophys. J.*, **286**, 503–508.

Readhead, A. C. S., Walker, R. C., Pearson, T. J., and Cohen, M. H. (1980), "Mapping radio sources with uncalibrated visibility data", *Nature*, **285**, 137–140.

Readhead, A. C. S. and Wilkinson, P. N. (1978), "The mapping of compact radio sources from VLBI data", *Astrophys. J.*, **223**, 25–36.

Roberts, D. H., Gabuzda, D. C., and Wardle, J. F. C. (1987), "Linear polarization structure of the BL Lacertae object OJ 287 at milliarcsecond resolution", *Astrophys. J*, **323**, 536–542.

Roberts, D. H., Potash, R. I., Wardle, J. F. C., Rogers, A. E. E., and Burke, B. F. (1984), "Milliarcsecond polarization measurements", in *VLBI and Compact Radio Sources*, R. Fanti, K. Kellermann, and G. Setti, Eds., IAU Symposium No. 110, D. Reidel (Dordrecht, Holland), pp. 35–38.

Roberts, D. H. and Wardle, J. F. C. (1986), "Polarization distributions in compact radio sources", in *Quasars*, G. Swarup and V. K. Kapahi, Eds., IAU Symposium No. 119, D. Reidel (Dordrecht, Holland), pp. 141–147.

Schwab, F. R. and Cotton, W. D. (1983), "Global fringe search techniques for VLBI", *Astron. J.*, **88**, 688–694.

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

20. VERY LONG BASELINE INTERFEROMETRY II: THE TECHNIQUES OF SPECTRAL LINE VLBI

PHILIP J. DIAMOND

1. INTRODUCTION

This lecture discusses the techniques of spectral line VLBI. Spectral line VLBI is a generalization, and a natural extension, of continuum VLBI. Many techniques are common to continuum and line analysis; but it is the the important differences that are the subject of this lecture.

Molecular masers are the principal type of object observed with spectral line VLBI. At present, four different molecular masers have been detected with VLBI: OH, H_2O, SiO and CH_3OH. These masers are found in regions of star formation or in the circumstellar envelopes of late-type stars. OH and H_2O masers have also been detected in other galaxies using spectral line VLBI.

2. SPECTRAL LINE VLBI OBSERVATIONS AND CORRELATION

This section describes the acquisition of spectral line VLBI data and discusses its correlation in some detail. Later sections describe the data calibration and the methods used to obtain images of line sources. Most of the lecture is concerned with the use of the MkII VLBI system. I shall occasionally mention the MkIII system and shall explicitly note any differences between the two.

2.1. Data acquisition.

The VLBI observation of a spectral line source is almost identical to that of a continuum source (described in Lecture 19). There are, however, two important differences that reflect the nature of the objects observed.

When observing line sources it is usually important to obtain as high a frequency resolution as possible. Since correlators are limited by the number of channels built into the hardware, the normal way to increase frequency resolution is to observe with a narrower bandwidth. Most VLBI stations are equipped with a standard set of baseband filters that increase in binary steps from 62.5 kHz to 2.0 MHz. Choice of the appropriate filter to use depends on the velocity width of the source being observed. It is possible to maintain a high frequency resolution but still observe broad-band line sources by using either:

(a) frequency switching, in which the data being recorded on tape are obtained from a series of baseband video converters in a switching sequence; each video converter is set at a slightly different frequency to cover the whole velocity range of the source; or

(b) the MkIII VLBI system, in which the signals from up to fourteen separate video converters (each set to a slightly different frequency) can be written to tape simultaneously.

The second difference between line and continuum VLBI observing is that since we have only a limited bandwidth we must ensure that the line source being observed is present in our observing band. This means that the video converter frequencies (usually the final setting in the local-oscillator chain) are determined by (a) the rest-frequency of the maser transition being observed and (b) the velocity of the line source with respect to the LSR. Normally an observing frequency is calculated for each antenna and is kept constant when observing a particular source, so as to minimize the chance of human errors at the antenna. Fine adjustments necessary to correct for the Earth's motion are made after the data are correlated. These are discussed in Section 3.2.

2.2. Correlation.

Lecture 4 describes the operation of digital correlators in detail. I shall not replicate that discussion, but instead shall summarize the salient features of the correlation for spectral line work. The present NRAO MkII correlator has 576 channels that can be flexibly assigned to one of seven modes of operation (see Table 20–1). For normal spectral line VLBI it is run in mode '0'; that is, for each integration period the output of the correlator is a 192-channel cross-correlation function and two 96-channel autocorrelation functions. These correlation functions are written to tape for later calibration and analysis.

Table 20–1.
NRAO MkII Correlator Modes

Mode	Configuration
0	96 Ch AC A, 96 Ch AC B, 192 Ch XC A-B
1	288 Ch AC A, 288 Ch AC B
2	576 Ch AC A
3	192 Ch AC A, 192 Ch AC B, 192 Ch AC C
4	288 Ch XC A-B
5	128 Ch AC A, 128 Ch AC B, 128 Ch XC A-B
6	96 Ch XC A-B, 96 Ch XC A-C, 96 Ch XC B-C
7	64 Ch AC A, 64 Ch AC B, 64 Ch AC C
	64 Ch XC A-B, 64 Ch XC A-C, 64 Ch XC B-C

To understand the corrections that must be applied to the visibility data, it is useful to describe the signal path through the correlator in detail. Following Lecture 2 and Thompson, Moran and Swenson (1986) we can write down the phase response of an interferometer. If we assume that the system passes only the upper sideband, then the phases of the signals from the two antennas after mixing are

$$\begin{aligned} \phi_1(t) &= 2\pi(\nu - \nu_{\mathrm{LO}})t - \theta_1(t) - \Psi_1(\nu)\,, \\ \text{and} \quad \phi_2(t) &= 2\pi(\nu - \nu_{\mathrm{LO}})t - 2\pi\nu\tau_g - \theta_2(t) - \Psi_2(\nu)\,, \end{aligned} \tag{20–1}$$

where ν is the observing frequency, ν_{LO} is the signed sum of the frequencies of the local-oscillator chain, t is the time, $\theta_1(t)$ and $\theta_2(t)$ are the slowly varying terms that represent phase noise contributed by the frequency standards, $\Psi_1(\nu)$ and $\Psi_2(\nu)$ are the phase terms of the bandpass filters, and τ_g is the geometrical delay. The recorded signals have clock errors τ_{c1} and τ_{c2}, so the phases of recorded signals are

$$\begin{aligned} &\phi_1(t) = 2\pi(\nu - \nu_{\mathrm{LO}})(t - \tau_{c1}) - \theta_1(t) - \Psi_1(\nu)\,, \\ \text{and} \quad &\phi_2(t) = 2\pi(\nu - \nu_{\mathrm{LO}})(t - \tau_{c2}) - 2\pi\nu\tau_g - \theta_2(t) - \Psi_2(\nu)\,. \end{aligned} \tag{20–2}$$

During processing the processor advances the data stream from antenna 2 by an amount τ'_g, the estimate of τ_g, so

$$\phi_2(t) = 2\pi(\nu - \nu_{\mathrm{LO}})(t - \tau_{c2} + \tau'_g) - 2\pi\nu\tau_g - \theta_2 - \Psi_2(\nu)\,. \tag{20–3}$$

The output of the correlator is a cross-correlation function, or lag spectrum. The cross correlation functions (and autocorrelation functions) are corrected for the effects of the quantization loss (the Van Vleck correction—Lecture 4) as follows,

$$\rho(\tau) = \sin\frac{\pi\rho_c(\tau)}{2}\,, \tag{20–4}$$

where $\rho_c(\tau)$ is the raw cross-correlation coefficient and $\rho(\tau)$ is the corrected coefficient. The corrected cross-correlation functions are then Fourier transformed (with respect to lag, τ), to become cross-power spectra. The phase of one frequency channel (for frequency ν) at this point is

$$\phi_{12}(t) = \phi_1(t) - \phi_2(t)\,. \tag{20–5}$$

If we define

$$\begin{aligned} \Delta\theta(t) = \theta_2(t) - \theta_1(t)\,, \quad \Delta\Psi(\nu) = \Psi_2(\nu) - \Psi_1(\nu)\,, \\ \Delta\tau_c = \tau_{c2} - \tau_{c1}\,, \quad \text{and} \quad \tau'_g = \tau_g - \Delta\tau\,, \end{aligned} \tag{20–6}$$

where $\Delta\tau$ is the error in setting the delay (introduced because the processor tracks the geometrical delay in increments of 0.25 μsec), then

$$\phi_{12}(t) = 2\pi\nu_{\mathrm{LO}}\tau_g + \Delta\theta(t) + \Delta\Psi(\nu) + 2\pi(\nu - \nu_{\mathrm{LO}})\Delta\tau_c - 2\pi(\nu - \nu_{\mathrm{LO}})\Delta\tau\,. \tag{20–7}$$

As explained in Lecture 2, the fringes are then "stopped"—i.e., the fringe rate is reduced to nearly zero—by rotating the phases by an amount $2\pi\nu_{\mathrm{LO}}\tau_g$. The residual phase of the signal then becomes

$$\phi_{12}(t) = 2\pi\nu_{\mathrm{res}}t + \Delta\theta(t) + \Delta\Psi(\nu) + 2\pi(\nu - \nu_{\mathrm{LO}})\Delta\tau_c - 2\pi(\nu - \nu_{\mathrm{LO}})\Delta\tau\,, \tag{20–8}$$

where $2\pi\nu_{\mathrm{res}}t$ is the difference between the fringe rotation frequency determined with perfect values of ν_{LO} and $\tau(t)$ and that derived from the (imperfect) values of the parameters adopted.

This formulation of the fringe phase is used since the five terms in Equation 20–8 are usually treated separately. The first term is the "geometrical" part of the residual fringe phase. It is nonzero if there are errors in the model used to calculate τ_g in the correlator. The second term, $\Delta\theta(t)$, is the phase noise of the frequency standards at the antennas; it is this term that limits the coherence time of the interferometer. The third term, $\Delta\Psi(\nu)$, is the difference in the phase response of the bandpass filters in the antennas. The fourth term describes the effect of clock differences at the stations. This clock difference, $\Delta\tau_c$, is one of the variables determined by fringe fitting. The fifth term has the same functional form as the fourth but is treated differently—it describes the effect of the digital delay tracking in the correlator. I have separated the last two because the $\Delta\tau_c$ term has only a slow time dependence whereas the $\Delta\tau$ term changes rapidly. I discuss its effects below.

Above I have described only the phase of the signal as it passes through the interferometer and correlator. To generalize this treatment, we should use the expression for the cross-power spectrum given by

$$S_{ij}(\nu,t) = A_i(\nu)A_j(\nu)g_i(\nu)g_j(\nu)e^{i\phi_{ij}(\nu,t)}\,, \tag{20–9}$$

where the A's are the amplitudes induced by the source signal and the g's are the amplitude responses of the bandpass filters at the antennas.

2.3. Digital delay tracking and the fractional bit shift error.

The digital nature of correlators implies that their output is not the perfect correlation function that we seek. The effects of the quantization loss, the fringe rotation loss and the fringe sideband rejection loss are described elsewhere (e.g., Lecture 4 and Thompson *et al.* 1986). There is a further loss because of the quantization of the estimated delay—e.g., in the NRAO MkII correlator τ_g' is set only every 0.25 μsec. Since the geometric delay is a continuously varying function, the delay of the central channel of the correlator oscillates about the ideal continuous function and can be in error by up to ± 0.125 μsec. This effect is called the discrete delay step loss or the *fractional bit shift* error.

We can quantify this effect as follows: From Equation 20–8 we know that the phase introduced by the fractional bit shift error is

$$\phi_{\rm fbs}(\nu) = 2\pi\nu'\Delta\tau\,, \tag{20–10}$$

where I have defined $\nu' = \nu - \nu_{\rm LO}$. For a correlator in which delays are set at the Nyquist rate—i.e.,

$$\Delta\tau = \frac{1}{2\Delta\nu}\,, \tag{20–11}$$

where $\Delta\nu$ is the baseband bandwidth—the peak-to-peak phase error is obviously

$$(\phi_{\rm fbs}(\nu))_{\rm pp} = \frac{\pi\nu'}{\Delta\nu}\,, \tag{20–12}$$

Figure 20–1. The fractional bit shift effect. Case **(a)** applies when the fringe rotator corrects the phase for zero baseband frequency (e.g., the NRAO MkII correlator), and case **(b)** applies when the fringe rotator also inserts a $\frac{\pi}{2}$ phase shift when the delay changes by one Nyquist sample (e.g., the CIT and MkIII correlators). The top plots show the phase versus time at baseband frequency ν'. The middle plots show the phase across the baseband at three different times denoted by 1, 2 and 3. The bottom plots show the averaged amplitude across the baseband. [Plot taken from Thompson, Moran and Swenson (1986).]

so that at the high-frequency edge of the band the phase error oscillates between $\pm 90°$ (see Fig. 20–1).

Because, over a period of time, the delay in the correlator is constant while the true delay is changed by the rotation of the Earth, the phase error introduced by the fractional bit shift has a sawtooth pattern as a function of time (see Fig. 20–1). The frequency of the sawtooth is directly proportional to the fringe frequency (i.e., the delay rate) and has a maximum value of

$$\nu_{\text{fbs}} = \frac{2\Delta\nu D\omega_e}{c}\,, \tag{20–13}$$

where D is the baseline length, ω_e is the angular velocity of Earth's rotation in rad sec^{-1} and c is the velocity of light. ν_{fbs} is in units of delay steps per second. If this effect is not corrected for and the data are averaged for a time much greater than $1/\nu_{\text{fbs}}$, then the phase at any frequency is uniformly distributed over $\phi_{\text{fbs}}(\nu)$. The amplitude loss as a function of baseband frequency is

$$L(\nu') = \frac{\sin(\phi_{\text{fbs}}(\nu)/2)}{\phi_{\text{fbs}}(\nu)/2}\,. \tag{20–14}$$

The corresponding net loss in signal-to-noise over a bandwidth of $\Delta\nu$ is 0.873, but at the high-frequency edge of the band the loss factor is 0.64. As an example, consider a typical spectral line VLBI experiment observed with the EVN in which the maximum baseline length is $\sim$ 1000 km and the minimum $\sim$ 300 km, and the observing bandwidth is 250 kHz. From Equation 20–12 we see that

$$\phi_{\text{pp}} = \frac{0.25\pi}{2.0} = 22\overset{\circ}{.}5\,. \tag{20–15}$$

The maximum frequency of the sawtooth pattern (Eq. 20–13) is $\sim$ 0.25 Hz; that is, the delay is updated every 8 seconds on the longest baseline but only every 24 seconds on the shortest. The net effect is to introduce a time-varying phase error that is most extreme at the high frequency edge of the band. For a typical continuum VLBI experiment on the same network, the peak-to-peak phase error at the high-frequency end of the band would be 180°, but the delay update rate would be eight times faster than that estimated above. The phase error would therefore be smeared out with a small amount of averaging, but the amplitude (and signal-to-noise ratio) would still be reduced.

To correct the data for this fractional bit error we must multiply the cross-power spectrum by $e^{2\pi i\nu'\Delta\tau}$. Previously, the correction for the fractional bit shift error was made in the correlator as the Fourier transform to the frequency domain was done. However, we now dump correlation functions to tape and make the full correction later.[1]

3. Calibration of Spectral Line VLBI data

The theory of calibrating spectral line VLBI data is the same as that of normal interferometric calibration. The practical details differ considerably, however, so this section describes the process as it is now carried out. The normal order of calibration is (1) bandpass calibration, (2) Doppler correction, (3) antenna gain estimation, (4) editing, and (5) fringe fitting (and phase referencing). I shall discuss each of these in turn.

[1] The correction can be made within AIPS using the task 'VLBIN'.

3.1. Bandpass calibration.

Bandpass calibration is the process of determining the shape of the bandpass filters for an antenna and correcting the data for the effects of these shapes. The bandpass response function, $g_i(\nu)g_j(\nu)\exp(i\Delta\Psi(\nu))$, is usually determined by observing a flat-spectrum continuum source.

Normally, for VLBI, we determine only the amplitude correction and assume that the phase response of the filters $\Delta\Psi(\nu)$ is flat and zero. It is possible to determine the phase response of the filters after fringe fitting, but this is beyond the scope of this lecture.

The bandpass calibration of VLBI spectral line data is presently accomplished by defining two quantities: $S_{\mathrm{off}_j}(\nu)$ is the off-source power spectrum for antenna j (i.e., the Fourier transform of the off-source autocorrelation function corrected for the effects of clipping), and $S_{\mathrm{on}_j}(\nu)$ is the on-source power spectrum. The bandpass-corrected cross-power spectra are then given by

$$S_{ij}(\nu) = \frac{S_{ij}^{\mathrm{raw}}(\nu)}{\sqrt{S_{\mathrm{off}_i}(\nu)S_{\mathrm{off}_j}(\nu)}}\,, \tag{20–16}$$

where $S_{ij}^{\mathrm{raw}}(\nu)$ is the Fourier transform of the raw cross-correlation function for baseline i-j. The $S_{\mathrm{off}}(\nu)$ are the clipped and normalized equivalents of the $g(\nu)$ mentioned above.

The corrected total-power spectrum for antenna i is given by

$$S_i(\nu) = \frac{S_{\mathrm{on}_i}(\nu) - S_{\mathrm{off}_i}(\nu)}{S_{\mathrm{off}_i}(\nu)}\,. \tag{20–17}$$

3.2. Doppler correction.

To simplify the experimental procedure at the antennas and to minimize the possibility of errors, the observing frequency is usually calculated before the experiment and set to a single value throughout the observation. Also, in present day VLBI systems the frequency of the final oscillator in the local-oscillator chain usually cannot be set to a precision better than 10 kHz. The net result is that the data after correlation all correspond to the same frequency. This causes an error, because the frequency at which a source of given LSR velocity should be observed changes as the Earth rotates. This effect makes a line at a given velocity wander, over time, backward and forward through the spectrum.

This Doppler effect can be corrected using the Shift Theorem for Fourier transforms. The procedure is to Fourier transform the observed spectra to correlation functions, multiply the correlation functions by $e^{i2\pi\Delta\nu\tau}$, where $\Delta\nu$ is the desired frequency shift, and then Fourier transform the correlation functions back to spectra. All the data will then have a constant velocity instead of constant frequency, and the position of a given line feature within the spectrum will be constant with time.

3.3. Antenna gain estimation.

Lecture 19 explained that the normal way to calibrate continuum VLBI data uses the formula

$$S_{ij} = \widetilde{A}_{ij} b \sqrt{\frac{T_{\mathrm{sys}_i} T_{\mathrm{sys}_j}}{K_i K_j}}\,, \tag{20–18}$$

where S_{ij} is the calibrated, correlated flux density, $\widetilde{A}_{ij}$ is the raw correlation coefficient, b is the quantization correction factor, the K_i are the antenna sensitivities in K Jy^{-1}, and the T_{sys_i} are the on-source system temperatures in Kelvins; the latter two quantities are obtained at each telescope during an observing session.

For spectral line VLBI we can use a much more powerful method of antenna calibration. We know that the total-power spectrum of a source, $\widetilde{S}_i(\nu)$, observed at a given antenna is related to the true, calibrated total-power spectrum $S(\nu)$ by the relation

$$\widetilde{S}_i(\nu) = Q_i S(\nu)\,. \tag{20–19}$$

The factor Q_i is a "quality factor" that is directly proportional to the effective collecting area of the antenna, divided by the system temperature; it also includes all the effects that can affect the cross-power data, such as pointing errors or atmospheric and receiver fluctuations. Therefore we need only one good, calibrated total-power spectrum (so in theory we need calibrate in the normal way only at one antenna during the observation), and then we can determine the quality factors as functions of time for each of the other antennas.

We determine the Q's by minimizing the sum of the squares of the quantity $\widetilde{S}_i(\nu) - Q_i S(\nu)$ over the spectrum; i.e., we make a least-squares fit of each antenna's total-power spectra to the good, calibrated total-power spectrum. This procedure then provides us with a so-called "gain table"; i.e., a list of Q's as a function of time for each antenna. These quality factors are then used for the final calibration of the bandpass-corrected cross-power spectra using the relation

$$S_{ij}(\nu) = \widetilde{A}_{ij}(\nu) \frac{b}{\sqrt{Q_i Q_j}}\,. \tag{20–20}$$

For the NRAO MkII correlator, the quantization correction factor b is 1.28 (i.e., it includes only the effects of the digital fringe rotator, since the Van Vleck correction is made earlier).

This method of antenna gain calibration can be used only in the presence of a strong spectral line, such as a maser line or a deep H I absorption line. It can also be a powerful method of calibrating high-frequency continuum data. For example, if you wish to make a continuum observation at a frequency at which most antennas have poor pointing and low efficiencies (e.g., 43 GHz), and if there is a strong spectral line source close to the program source, you can observe the line source every few minutes to calibrate the gains and can calibrate the continuum data by interpolation.

Figure 20–2. Fringe fitting in action. The left-hand plot shows the complex cross-power spectrum for the source 3C 345 on the baseline Effelsberg–Jodrell Bank at 18 cm. The data have been coherently averaged for 5 minutes and no fringe fitting (i.e., calibration of fringe rate and delay residuals) has been done. The right-hand plot shows the same data after fringe fitting; all other parameters are the same. Notice how the phase slope has been removed (i.e., the residual delay has been corrected) and how the amplitude is much greater (i.e., a high fringe rate has been removed).

3.4. Editing.

Editing is the process of removing discrepant data. I shall not discuss this topic, as it has been dealt with in other lectures. I shall just point out that in editing line data we must be aware of the possibility of frequency-dependent interference; however I have not yet come across a VLBI case in which this has been important.

3.5. Fringe fitting.

Fringe fitting was described in detail in Lecture 19 (Sec. 4.1), so I shall not comment on it except to emphasize some practical aspects of fringe fitting to spectral line data. Basically, fringe fitting is the procedure used to determine the residual fringe rates and delays for the antennas in a VLBI experiment. The data are corrected for these residuals, effectively removing phase slopes in time and frequency. In spectral line VLBI it is difficult to determine the residual delay from the spectral sources, since the structure of the source changes as a function of frequency. Instead, we determine the residual fringe rates and delays from continuum sources which should be observed frequently during an experiment (typically every two hours or so). These residuals are interpolated into the spectral line data, thereby correcting the delay offsets and making a first-order correction to the fringe rates. Figure 20–2 shows the dramatic improvement in data quality obtained after fringe fitting.

There will be a residual fringe rate in the data for the line source since, unless you are lucky, the line source will be in a different part of the sky than the continuum calibrator. Therefore we select a strong channel (or group of channels) from the spectral line source and solve for its residual fringe rate in the same way as for the continuum sources. This residual fringe rate is then removed from all channels in the spectrum, dramatically slowing the phase variation with time.

The data should then be *phase referenced*—i.e., the phase of a point-like feature should be subtracted from all other features. Since any instrumental and atmospheric phase errors are usually common to all frequencies in the observing band, this process effectively removes these errors from the data. If the reference feature is resolved, the phase due to its structure will also be subtracted during phase-referencing, and this effect must be corrected for.

4. Spectral Line Imaging

Images of spectral line data can be obtained from visibility data in the same way as for continuum data, that is, by Fourier synthesis. In the general VLBI spectral line case, however, synthesis techniques can be prohibitively expensive—the sources may be spread over areas of sky four orders of magnitude larger than the synthesized beamwidth θ_{HPBW}. This is especially true of the water masers at 1.35 cm. The resolution of an intercontinental experiment at this wavelength is about 0.2 milli-arcseconds, and the masers are usually spread over 2 arcseconds, and sometimes as much as 30 arcseconds. An image made with normal imaging parameters—3 points per beam covering twice the field containing the source—is out of the question with any computer. It would take over 3.6 Gbytes just to store the image of one channel in the 2 arcsecond case! To compound the problem, the masers often have complex spectra that require many channels to be imaged.

Figures 20–3 and 20–4 show an example of such a complex maser source. Figure 20–3 shows the spectrum. The maser region consists of hundreds of separate features, each a km s^{-1} or two wide and spread over 400 km s^{-1}. The spectrum shown contains about 1500 frequency channels, each of which must be mapped. Figure 20–4 shows the layout of the source derived by *fringe rate mapping* (see below). The features are spread over more than two arcseconds, while the resolution of the experiment, θ_{HPBW}, is less than one milli-arcsecond. This display shows only the rough distribution and intensities of features; blow-ups of individual regions are needed to show all the information.

One consequence of the large fields-of-view, measured in beams, is that the data cannot be averaged for long, for the reasons discussed in Lecture 13. Typical water maser observations are limited to one or two seconds of integration. Orion (30 arcseconds in extent) is limited to 0.2 seconds of averaging, and less might be preferred. These short averaging times, combined with the need to observe many spectral channels, lead to large data sets. One reason that so few spectral line VLBI experiments have been done, and that some of those that have been done are taking so long to reduce, is just that the data volume is overwhelming.

Figure 20–3. A spectrum of a complex water maser source (W49 — Walker, Matsakis, and Garcia-Barreto 1982). There are about 1500 spectral channels in the spectrum. The solid line shows the total power received by a single antenna, while the dots show the portion of the total power that was successfully mapped in a 3-station experiment using fringe rate mapping techniques. There is clearly a lot of missing flux density in the maps, indicating that much could be learned with proper synthesis observations, but, as discussed in the text, the volume of data in such observations is extreme.

Clearly, clever methods must be found to make the imaging problem reasonable. The sources often consist of a few point-components at each frequency. Calculating and storing 3.6 Gbytes, just to determine the parameters of a handful of separate features, is inefficient. The method that has been used most is called *fringe rate mapping.* Not only does it avoid using a grid, but also it uses only the rate of change of phase rather than the phase itself, so it is not sensitive to delay offsets. If the amplitudes are at all stable it can be done on nearly uncalibrated data—only the phase referencing step is critical.

The method is based on the fact that the fringe rate or rate of change of phase, $\dot{\phi}$, for a feature offset from the phase center by the angular coordinates Δl and Δm is given by

$$\dot{\phi}(\nu) = 2\pi \left(\dot{u}\Delta l + \dot{v}\Delta m \right) , \tag{20–21}$$

where u and v are the projected baseline components in wavelengths. The phase center is established during the phase referencing procedure and will be given by the centroid of the emission of the channels used to determine the residual fringe rate. The method is based on the fact that there is a phase offset between any two features that are separated on the sky, but also that this phase offset changes with time. The rate of change is the relative fringe rate; it can be as high as 0.2 Hz per arcsecond at 1.35 cm on intercontinental baselines.

The first post-calibration step of fringe rate mapping is to calculate fringe rate spectra for each channel. A fringe rate spectrum may be obtained by performing a Fourier transform on a time sequence of amplitude and phase data. Ordinarily such a spectrum is generated for each frequency channel in the cross-power spectrum that contains significant flux. The choice of the interval of data over which to calculate each fringe rate spectrum is based on a tradeoff.

Figure 20–4. The fringe rate map of the water maser source whose spectrum is shown in Figure 20–3. The main concentration of features is spread over 2 arcseconds, with a few features over a larger region. The sizes of the symbols represent the flux densities of the features. The resolution is better than a milli-arcsecond, so it is difficult to display the full spatial dynamic range in one image. The large ratio of source area to beamwidth θ_{HPBW} leads to unreasonably large images if straightforward synthesis imaging is used. Either a gridless method such as fringe rate mapping or model fitting must be used, or the imaging must be restricted to small fields around individual features.

Figure 20–5. The lower plot is a fringe rate spectrum of one velocity channel for the source in Figure 20–3. There are four peaks, each corresponding to a separate feature on the sky. Each peak constrains its corresponding feature to lie along a line on the sky. The upper plot shows such lines from many scans. The peaks in the lower plot and their corresponding lines in the upper plot are labeled A–D. There are clearly four separate features at the velocity of these data, including one (corresponding to line D) that is sufficiently far from the phase center that smearing of the fringe rate peaks will prevent derivation of an accurate position. The window in which reasonable positions can be found is about 0.5 arcseconds in R.A. and 2.0 arcseconds in declination for this low-declination source with 20-min integrations. The window can be moved by shifting the phase center of the data. [Taken from Walker (1981).]

Long integrations give higher signal-to-noise ratio and finer fringe rate resolution. However, relative fringe rates change with time, since $\dot{u}$ and $\dot{v}$ are not zero, so the fringe rate peaks will be smeared if the integration time is too long. Typically, times ranging from a few minutes to an hour are used.

Next, each fringe rate spectrum is examined for peaks, and the parameters of each peak are extracted. An automatic program is available to do this. Each fringe rate peak constrains the position of the associated feature to a line on the sky. The slope of the line is $\tan^{-1}(\dot{v}/\dot{u})$ so, as the projected baseline changes, the slopes of the lines change. The lines for all peaks from one fringe rate spectrum are parallel. Since $\dot{u}$ and $\dot{v}$ change with time, it is possible to find the position of the emission by finding the point at which the lines corresponding to different scans intersect. If the lines from all fringe rate spectra for a channel are plotted, it is easy to pick out by eye the places where many lines intersect. Figure 20–5 shows an example of a fringe rate spectrum and of the plotted lines corresponding to all of the fringe rate spectra for a single channel from a three-station experiment. The trick is to select automatically all the peaks that correspond to one feature and use them in a least-squares fit for the position of the feature. This selection is complicated by several effects: Features may overlap in fringe rate space; sometimes features are not seen in all fringe rate spectra, for noise or dynamic range reasons; sometimes the automatic peak-finding routine finds false peaks. A program has been written that tries to disentangle all this, to select the fringe rate peaks that correspond to each feature. It should be checked by plotting the lines and examining the plot by eye. More sophisticated mathematical techniques might be used in future programs.

The accuracy of positions found with fringe rate mapping is a few times worse than that achievable with synthesis imaging. The sensitivity of relative fringe rate to a position offset is much lower than the sensitivity of relative phase. The accuracy depends both on the sensitivity to position offsets and on the accuracy with which the location of each fringe rate peak can be determined.

An alternate mapping technique is to use a low-resolution synthesis image or a fringe rate map to identify the locations of features, and then to make images of small fields around each one.[2] This method will give full resolution, in case there is any interesting structure in the individual features. However, it will miss any features not found in the low-resolution maps, which are likely to be based on smaller amounts of data and to have lower dynamic range.

It would be possible to devise a gridless method based on relative phase that would obtain much higher position sensitivity than the fringe rate method. However, it would require either data in which the phase slopes have been removed or a method for fitting for delays along with position. Perhaps the nastiest problem would be that the 2π ambiguities in phase would have to be resolved. Fringe rates are not subject to such ambiguities.

When mapping line sources that cover only a small area of sky (e.g., stellar masers) it is possible to use the synthesis technique alone, since the ratio of map size to beam size is much smaller than in the case discussed above, and the storage requirements are not ridiculous.

Once the maps are made, we are faced with all the usual problems of how

[2] The AIPS program 'MX' is well-suited to this purpose.

to display spectral line data. But that is yet another extensive topic.

References

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

Walker, R. C. (1981), "The multiple-point fringe-rate method of mapping spectral-line VLBI sources with application to H_2O masers in W3–IRS5 and W3(OH)", *Astron. J.*, **86**, 1323–1331.

Walker, R. C., Matsakis, D. N., and Garcia-Barreto, J. A. (1982), "H_2O masers in W49N. I. Maps", *Astrophys. J.*, **255**, 128–142.

21. SOLAR IMAGING WITH A SYNTHESIS TELESCOPE

Tim Bastian

1. Introduction

In this lecture I consider the the problem of imaging the Sun with a synthesis radio telescope. Until the advent of large synthesis telescopes such as the Westerbork Synthesis Radio Telescope (WSRT) or the Very Large Array (VLA), most solar imaging experiments were confined to constructing the *average* distribution of radio brightness across the solar disk using low-resolution fan beams (e.g., Christiansen and Warburton 1955) or to imaging the instantaneous brightness distribution using low-resolution pencil beams. Examples include the **T**-array at Fleurs and the circular array at Culgoora, both in Australia. It is difficult to image the Sun using the well-developed principles of Earth-rotation aperture synthesis with a small number of antennas because solar radio emission is highly dynamic—the u-v plane must be sampled rapidly to obtain a useful result.

When the WSRT became available, however, many more baselines could be sampled instantaneously, and high-resolution solar imaging became possible. Bregman and Felli (1976) made the necessary modifications to the WSRT for solar observing and produced the first high-resolution images of the Sun at a wavelength of 21 cm. Since then, the WSRT has been used on several occasions to image the quiet Sun in two dimensions at both 6 cm and 21 cm (e.g., Kundu *et al.* 1979) and to produce one-dimensional images of microwave burst sources (Kattenberg and Palagi 1983). Even so, the solar imaging capabilities of the WSRT are limited because it is a one-dimensional array, and many hours are required to produce a two-dimensional image.

The Clark Lake Radio Observatory (CLRO) was also an important pioneering synthesis telescope. Not only could it image the Sun routinely, but it was one of the few low-frequency synthesis telescopes; most others operate at centimeter wavelengths. Unfortunately, the CLRO is no longer operational. The East–West array at Nançay, France, has also been used to make synthesis images of the quiet Sun at meter wavelengths (169 MHz).

The VLA is by far the most powerful synthesis telescope for solar imaging. Its power lies in its ability to make high-resolution two-dimensional images of the Sun quickly at many frequencies. While the VLA is far from being the solar observers' ideal,[1] it is nevertheless the best available for solar observing. My discussion will therefore be slanted toward solar imaging as practiced at the VLA, and the problems that must be confronted there.

[1] "The VLA was not *designed* to observe the Sun, for God's sake. Why do you keep looking at it?"—*Anon.*

2. Solar Imaging with the VLA

2.1. 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 (Table 21–1). Two fundamental hardware modifications are therefore necessary for solar observing. I discuss each of these in turn.

Table 21–1.
Solar Observing Parameters at the VLA

Band (cm)	$T_{\rm sys}$ (K)	$T_{\rm ant}$(QS) ($\times 10^3$ K)	$T_{\rm ant}$(AR) ($\times 10^3$ K)	$T_{\rm ant}$(burst) ($\times 10^3$ K)	$T_{\rm ant}$(AR)/$T_{\rm sys}$ (dB)
90	208	18	42	750	23
20	67	42	100	”	32
6	50	17	75	”	32
3.6	35	11	48	”	31
2	130	8	30	”	24
1.3	270	6	20	”	19

2.1.1. Gain reduction.

The enormous increase in system temperature which occurs when observing the Sun is far greater than the electronics can properly handle. Specifically, the automatic level control loops, or ‘ALC’s, are unable to reduce the overall system gain sufficiently to ensure a constant power to the correlators. Thus, 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. It turns out that the first amplifiers have plenty of dynamic range and can handle all expected system temperature increases from the Sun, so the attenuation is inserted after these amplifiers. The choice of the value for the insertion loss introduced by these switched attenuators is driven by constraints imposed by the ALC loops, and by the microwave-emitting properties of the quiet and active Sun.

ALC loops are discussed in Lecture 4. Recall that from an electrical engineering standpoint, it is much easier to maintain constant signal-power inputs to the correlator than it is to maintain a constant gain for each antenna. ALC loops use variable attenuators to adjust the gains so that the inputs to the correlator have constant power. The ALC loops show a significant phase variation with gain (Fig. 21–1). For reasons discussed further in Section 2.2, it is desirable to minimize phase differences introduced by the ALC loop between calibrator and solar scans. This is done by “tuning” the attenuators to values that most nearly satisfy this requirement.

The choice of insertion loss must, however, be a compromise, because of the radio-emitting properties of the Sun. Radio emission from the Sun may be

Figure 21–1. Phase shift introduced by the ALC attenuators as a function of insertion loss for the two types of attenuator used at the VLA.

loosely divided into three categories: (i) an omnipresent *background* component that varies at the leisurely pace of the eleven-year sunspot cycle; (ii) a *slowly varying* component associated with active regions on the solar disk, which varies on time scales from hours to days; and (iii) a *transient* component associated with solar flare activity, which varies on time scales from less than a second to hours. Table 21–1 shows antenna temperatures representative of the quiet Sun, of a strong ($\sim$ 100 sfu)[2] active region[3] in the beam at 21 cm, and of a strong ($\sim$ 1000 sfu) microwave burst in the beam. Table 21–1 shows that the insertion loss introduced by the ALC loop must vary considerably in response to solar conditions.

The gain-reduction scheme implemented at the VLA is based on values of T_{ant} similar to those in column 4 of Table 21–1; these represent the best compromise between active- and quiet-Sun observing. For all bands, *phase-constant* 20 dB switchable attenuators are used. As shown in the last column of Table 21–1, however, the gain reductions required at 20, 6 and 3.6 cm are substantially greater than those required at 90, 2, and 1.3 cm. Hence, at 20, 6, and 3.6 cm extra attenuators are used; at present, 12 dB of extra attenuation are inserted at 20 cm and 9 dB of extra attenuation are inserted at 3.6 and 6 cm.

2.1.2. Measurement of T_{sys}.

Since, under the action of the ALC loop, the antenna gain is allowed to

[2] One sfu (solar flux unit) = 10^4 Janskys = 10^{-22} W m^{-2} Hz^{-1}.

[3] A typical active region spectrum is assumed except at 90 cm where several Type I storms are assumed.

vary, the antenna gains must be monitored. In normal observing, 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 $T_{\rm cal} \approx 0.1 T_{\rm sys}$. Separate noise sources are provided for each band and for each of the two orthogonal senses of polarization. The antenna gains are calibrated by measuring two voltages in the front end of each antenna: the *synchronous-detector voltage* and the *total power voltage*. The synchronous-detector voltage $V_{\rm SD}$ measures the *difference* between the total power levels when the noise source is on and when it is off. The total power voltage $V_{\rm TP}$ measures the power level when the noise source is off. The former is, of course, proportional to $T_{\rm cal}$, the injected noise, and the channel gain; the latter is proportional to $T_{\rm sys}$ and the channel gain. It follows that

$$T_{\rm sys} \propto \frac{V_{\rm TP}}{V_{\rm SD}}\, T_{\rm cal}\,. \tag{21–1}$$

As the gain is inversely proportional to $T_{\rm sys}$, the antenna gains are corrected 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 noise sources are employed. However, because of the expense of retrofitting all twenty-eight VLA antennas with high-temperature noise sources (*solar calibration sources*), only four VLA antennas are provided with solar calibration noise sources at most frequencies. The exceptions are at at 20 cm, where five antennas have been so outfitted, and 3.6 cm, where all antennas have been.

The solar calibration noise sources have been adjusted so that $T_{\rm cal}$ is $\sim 10\%$ of the system temperature that results from quiet-Sun conditions. A stronger solar calibration signal could cause problems with the operation of the correlator under some circumstances (Archer 1979). On the other hand, since the minimum calibration signal that may be reliably detected is of order 0.5% of $T_{\rm sys}$, the useful dynamic range 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 an electronic component—with this criterion, the dynamic range of the receiving system is roughly 20 dB, or a factor of 100 over quiet-Sun conditions.

To put this in more concrete terms, the available dynamic range in the ALC loop can accommodate the increase in system temperature produced by most microwave bursts. It is found that the integral number of flares above a given amplitude is somewhat insensitive to frequency. For example, Croom (1971) found that, during a 2.5 year period near solar maximum, of 148 microwave bursts recorded at 19 GHz, sixty-five had maximum flux densities greater than 10^2 sfu, fifteen greater than 10^3 sfu, and two greater than 10^4 sfu. Similarly, at 2.8 GHz it was found (Kundu 1965) that 116 bursts had maxima greater than 10^2 sfu and six were greater than 10^3 sfu. Hence, using the 6-cm antenna temperature listed in column 4 of Table 21–1, one expects to be able to observe flares with total power amplitudes of up to 900 sfu with relatively high accuracy, and up to 4500 sfu with less accuracy.

2.2. Solar data calibration.

2.2.1. Amplitude calibration.

The calibration of visibility amplitudes is a multi-stage process. As described above, a rough calibration is accomplished on-line by referencing the input signal to a noise source of known amplitude. Normally, the amplitude calibration is further refined by referencing the source *and* a calibrator source to the same noise source, where the flux density of the calibrator is presumed to be accurately known (Lecture 5). Self-calibration corrects most of the remaining errors. When observing the Sun, there are *no* calibrator sources available that are strong enough that both the Sun and the calibrator can be referenced to the same noise source. Thus, as described in Section 2.1.2, two noise sources must be used; the normal ones and the special solar calibration noise generators.

In practice, two factors limit the accuracy of the solar amplitude calibration. First, the gains of the antennas that are not outfitted with solar calibration noise sources 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 calibration signals are known. Since there is no celestial source of known flux density bright enough to serve as a reference for the solar calibration signals, they must be measured individually in the field. The accuracy of this measurement is of order 20% (P. Lilie, private communication).

With the above difficulties in mind, the amplitudes are calibrated as follows: $T_{\rm sys}$ is measured during a solar observation for a small subset of antennas. On the basis of consistency of the measured $T_{\rm sys}$ between these antennas, one identifies several antennas as reference antennas. Using these antennas, an average $T_{\rm sys}$ correction is formed as a function of time and is applied to the antennas that lack solar calibration noise generators.

2.2.2. Phase calibration.

Phase calibration of solar data is more straightforward, in the sense that the phase solution is transferred directly from the calibrator source to the solar visibilities. Recall that the 20 dB switched attenuators described in Section 2.1.1 are phase-constant. In other words, they have been designed to introduce a negligible phase change when switched in or out. While the second-stage attenuators used for observations at 3.6, 6, and 20 cm are not phase-constant, they are normally not switched—when observing the Sun at 3.6, 6, and 20 cm, the second-stage attenuators remain in place for both calibrator and source scans. Hence, the accuracy of the phase transfer depends primarily on the extent to

Figure 21–2. Phase-calibration transfer errors for: **(a)** quiet-Sun conditions; **(b)** 20 dB above the quiet Sun.

which the ALC insertion loss is the same on both the calibrator source and the Sun. This is roughly true in most circumstances, but is far from true in a few exceptional cases.

Figure 21–2a summarizes the expected range of phase-calibration transfer errors for quiet-Sun observations, and the nominal $T_{\rm sys}$, as a function of the ALC-loop insertion loss. Figure 21–2b shows the same quantities for a source 20 dB above quiet Sun. So long as the ALC attenuator remains within ± 3 dB of its nominal operating point (-12 dB), phase-calibration transfer errors are less than 3° for quiet-Sun observations. For a source 20 dB above quiet Sun, the errors increase to 10°–15°.

3. Topics in Solar Synthesis Imaging (or: "Why Does My Solar Image Look So Awful?")

Solar observers are often puzzled and unhappy when they find that their images are limited in dynamic range from a few times 10 : 1 to a few times 100 : 1. It is a sad fact that images of the Sun made using the VLA will *never* approach the quality of those made of other sources. The reasons for this are many, and merit some discussion.

3.1. Sensitivity.

One might assume that when observing the Sun, like any other source, the dynamic ranges of one's images are limited first by calibration errors, but ultimately by $T_{\rm sys}$, the integration time $\tau_{\rm av}$, and the bandwidth $\Delta\nu$. As was shown in Lecture 7, when observing a strong fully resolved source where $T_{\rm ant} \gg T_{\rm sys}$, the sensitivity of a simple correlation interferometer is given by

$$\Delta S = \frac{\sqrt{2} k_B T_{\rm ant}}{\eta A \sqrt{\tau_{\rm av} \Delta\nu}}, \tag{21–2}$$

where ΔS is the r.m.s. variation in the flux S, k_B is Boltzmann's constant, η is the aperture efficiency and A is the area (each presumed identical for both antennas). When observing the Sun, the situation is complicated by the additional contribution to the noise produced by the correlated flux S_C on each baseline. Under many circumstances, the correlated flux is negligible compared to the total. When this is not the case (e.g., for the shortest baselines in compact antenna configurations or during a strong, compact microwave burst) one can no longer ignore the contribution to the noise due to the correlated flux on each baseline. Lecture 23 describes the properties of the noise in images of strong sources in detail. To provide rough estimates of the instrumental sensitivity, I assume that the Sun is strongly resolved so that $S_C \ll S$. Assuming $\tau_{\rm av} = 10$ sec and $\Delta\nu = 12.5$ MHz, and using the representative antenna listed in Table 21–1, I derive the VLA single-multiplier correlation interferometer sensitivities listed in Table 21–2.

Table 21–2.
VLA Sensitivities For Solar Observing
(Single Correlator, 12.5 MHz Bandwidth, 10 sec)

Band	$\Delta S_{\rm QS}$(Jy)	$\Delta S_{\rm AR}$(Jy)	$\Delta S_{\rm burst}$(Jy)
90 cm	21	49	875
20 cm	39	93	700
6 cm	12	54	540
3.6 cm	8	34	530
2 cm	7	27	670
1.3 cm	6	21	770

A striking example of the dependence of instrument sensitivity on $T_{\rm ant}$ is given by Gary and Hurford (1987), who used a partial eclipse of the Sun to image an active region. The method is illustrated in Figure 21–3a.

The lunar limb provides a moving knife edge; because the VLA uses a finite integration time (10 sec in this case), the strip of the Sun revealed in this time is weighted by a ramp function. If two successive integrations are differenced, the resulting visibilities are due solely to the strip of the Sun's brightness revealed during this time, weighted by a triangular window function. By using two subarrays, Gary and Hurford were able to image the Sun at 6 and 20 cm over the

Figure 21–3. A schematic illustration of the method used by Gary and Hurford (1987) to image the Sun during a solar eclipse. The motion of the lunar limb across the Sun acts as a knife edge. **(a)** shows the ramp profile caused by the Moon's motion during two successive integrations; **(b)** the window which results when the two ramps are differenced; **(c)** the point spread function (or beam) for normal observations; and **(d)** the effective point spread function resulting from the convolution of (b) with (c). (Reprinted with permission from the *Astrophysical Journal.*)

time required for the Moon to transit the area of interest; the resulting 6-cm image is shown in Figure 21–4. A gradient in the noise from West to East due to the roughly linear increase in T_{ant} as the Moon uncovered the Sun is readily apparent.

Unlike the situation in this example, noise fluctuations do *not* usually limit the achieved dynamic range, as I explain below.

3.2. The problem of calibration.

Section 2.1 described the hardware modifications to the VLA that make it possible to observe the Sun. The nature of the VLA electronics design made compromise unavoidable. The price is paid in the relatively poor calibration of solar visibilities that results. The absolute calibration of solar data depends on the accuracy with which the solar calibration signals are measured, estimated to be about 20%. The quality of solar images is determined by the relative accuracy of the calibration of amplitudes and phases. Errors in the amplitude calibration arise because a global T_{sys} correction is applied to all antennas rather than a true antenna-based correction. The magnitude of this error is difficult to estimate for all bands. However, since all antennas have solar calibration noise generators at 3.6 cm, we may assume that the spread in T_{sys} at this wavelength is representative. A recent test of several 3.6-cm pointings showed a dispersion of 10% for pointings at disk center and $\gtrsim 15\%$ for a pointing near the solar limb.

Particular 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 $\approx 5°$ are to be expected routinely and transfer errors $\approx 10°$–$15°$ are to be expected when the ALC operates far from its nominal regime, e.g., during a solar flare.

Lecture 16 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 $\sqrt{M}N/x$ 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 $\epsilon \approx 20\%$ and $\phi \approx 10°$ for all antennas then $D \approx 100\sqrt{M/2}$. 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.

3.3. The problem of confusion.

For many solar observing programs, the single most important factor determining the dynamic range is *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 24) 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 and 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

$$\sigma_c = \rho\sqrt{\int B^2(\Omega)\,d\Omega \int S^2 n(S)\,dS}\,, \tag{21–3}$$

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.

It is unclear whether it is possible or even appropriate to represent the Sun's radio brightness distribution as a distribution of point sources, $n(S)$. However,

Figure 21–4. The 6-cm map obtained by Gary and Hurford during a solar eclipse. The contour levels are spaced at intervals of 10% of the peak brightness temperature $T_B = 1.1 \times 10^6$ K. The succession of arcs indicates the position of the solar limb at 10 s intervals. Note the clear increase in noise from West to East due to the increase in antenna temperature as the Sun was exposed by the Moon.

recent work at 6 and 20 cm (e.g., Gary and Zirin 1988; Habbal and Harvey 1988) has shown that the radio brightness distribution of the quiet Sun is indeed composed of many discrete elements.

The work of Gary and Zirin at 6 cm (Fig. 21–4) shows that roughly 20% of the field of view contains discrete elements with brightnesses ranging from 10–43 $\times 10^3$ K above the background Sun ($\sim 18 \times 10^3$ K). Similarly, Habbal and Harvey have demonstrated the presence of 20-cm "coronal bright points". Like the discrete 6-cm features, the discrete 20-cm features range in brightness from 0.5 to 2.5 times the quiet-Sun brightness ($\sim 80 \times 10^3$ K). This situation is complicated by the fact that the radio brightness distribution on the Sun is composed of both discrete unresolved elements and structures that are strongly resolved. Further, the detailed distribution changes with time.

Nevertheless, to illustrate the consequences of confusion, I consider the effect of a random distribution of discrete sources within the primary beam with a dispersion in amplitude similar to that observed on the quiet Sun (Fig. 21–5a). A filling factor of $f = 15\%$ is assumed, although the result is not sensitive to f because $\sigma_c \propto \sqrt{n(S)}$. When convolved with a dirty beam corresponding to 6 hrs in the **C** configuration at 6 cm, the image in Figure 21–5b results. While the r.m.s. sidelobe level of the dirty beam is 0.5%, and one might therefore expect to achieve a dynamic range of 200 : 1, the confusion-limited image falls more than an order-of-magnitude short of this, the dynamic range being $\sim 15:1$. Admittedly, this treatment is quite crude and the problem needs to be treated in detail; yet it demonstrates the fundamental problem posed by confusion.

Figure 21–5. **(a)** A model 6-cm field containing a random distribution of point sources. The dispersion in source amplitudes is similar to that observed on the quiet Sun. A filling factor of 15% has been used. **(b)** The same field after tapering by the primary beam and convolution with the dirty beam resulting from 6 hrs of coverage by the VLA in the **C** configuration.

3.4. The problem of u-v coverage.

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 was discussed in Lectures 8 and 16. The problem is particularly acute for solar imaging because most of the power emitted by the Sun is in the largest angular scales.

Figure 21–6 gives an illustrative example. In Figure 21–6a, the Sun is represented as a disk with an "active region". Figure 21–6b shows the model brightness distribution that results when all spatial frequencies corresponding to those not measured by the VLA's **D** configuration have been filtered out, and an overall distortion to the brightness distribution is apparent. Hence, when measuring the brightnesses of quiescent solar phenomena (low brightness plage, coronal holes, filaments, etc.), it is essential to recover the shortest spatial frequencies.

This requirement may be fulfilled through the use of the powerful image reconstruction and deconvolution algorithms now available. Maximum entropy algorithms, described in detail in Lectures 8 and 15, provide a convenient way to introduce measurements made with more than one instrument to the image deconvolution problem. Hence, programs that depend on information contained

Figure 21–6. **(a)** A schematic representation of the solar disk with a small "active region". **(b)** The same after angular scales greater than 15′ have been filtered from the model, roughly corresponding to the Sun as viewed at 21 cm by the VLA in the **D** configuration.

in the shortest baselines may acquire it from a separate, single antenna observation of the Sun. Figure 21–7 gives an example in which the VLA and the 305-m reflector at Arecibo were used jointly to image the full disk of the Sun at 21 cm.

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. Until recently, long scans were relied upon for improved u-v coverage. Unfortunately, the Sun's radio brightness distribution *varies* in time due to (1) the Sun's rotation and (2) intrinsic source variability. The final problem that must be confronted is this variability.

3.5. The problem of 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.

Figure 21–7. An example of a full-disk solar image reconstruction at 21 cm for 3 September 1988. Here, low-resolution data from the Arecibo 305-m reflector were used to "fill in the hole" in the u-v plane caused by a minimum antenna spacing at the VLA. Bandwidth synthesis techniques were also used at the VLA to sample as much of the u-v plane as possible over the course of 4 hrs. The resulting image has a dynamic range of better than $300:1$, extremely high by solar standards. (Observers: T. Bastian, S. Habbal, R. Gonzalez, T. Cornwell, and G. Dulk.)

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. I return to this point below.

I now turn to the problem of intrinsic variability in the radio brightness distribution. Consider a discrete source at some location (l, m) 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 (l, m) is averaged down to $\tau S/T$. The net signal-to-noise ratio at (l, m) 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. For very strong, short-lived phenomena, the change in visibility can be detected and removed from each visibility. Unfortunately, many or most of these phenomena are not sufficiently strong to be detected in the visibility data.

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.

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. A promising approach to this problem is *bandwidth synthesis*. Since the baseline coordinates u, v, and w, are proportional to ν, the observed frequency, one can effectively change the baseline by changing ν.

As an example, Figure 21–8a shows the u-v coverage resulting from 6 hrs of observing at a single frequency within the 20-cm band in the **C** configuration. The r.m.s. sidelobe level of the dirty beam is roughly 0.5%. Figure 21–8b shows the coverage resulting from 1 hr of observing using five frequencies spanning a relative bandwidth of 20%. The resulting u-v coverage is quite good, resulting in a dirty beam with sidelobe levels comparable to that of the full 6-hr synthesis. Note that we rely on the non-coplanarity of the array for two-dimensional u-v coverage. Bandwidth synthesis with an East–West array is not useful because many hours are required to provide the necessary two-dimensional coverage. Implicit in the use of bandwidth synthesis, of course, is the assumption that one knows the spectrum of the emission over the bandwidth of interest reasonably well. On the other hand, a suitable choice of frequencies will provide baseline redundancy from which one might expect to extract information about the source spectrum.

The problem of source variability might look particularly grim for the solar phenomena that vary greatly in short times, i.e., microwave burst sources. Even bandwidth synthesis techniques are of little help here. Fortunately, microwave

Figure 21–8. **(a)** The u-v coverage resulting from 6 hrs of coverage at a single frequency in the 20-cm band in the **C** configuration of the VLA. **(b)** The u-v coverage resulting from 1 hr of coverage in the same configuration where 5 frequencies are used to sample a relative bandwidth of 20% of the 20-cm band.

burst sources tend to be compact, simple in morphology, and much brighter than the background Sun. Thus, a small number of visibilities, such as might obtained in a snapshot, are adequate to image them. Further, the simplicity and relative brightness of the sources lend themselves to self-calibration techniques. Thus, far from being the worst case, microwave burst sources are often a "best case" for solar work—in the sense that the limitations described above (calibration errors, confusion, u-v coverage) do not dominate.

Observations of burst sources may however be limited by the ALC loops, which were designed under the assumption that changes in T_{sys} over time would be slow. This is a fair assumption since changes in T_{sys} during normal observing are produced by changes in ground pickup, atmospheric emission, the galactic background, etc. The ALC loops were therefore designed with a 0.5 sec time constant, so they cannot respond to changes in the antenna gains faster than a few e-folding times. They cannot therefore reliably follow changes in the brightness of a burst source that occur faster than one or two seconds.

4. Solar Observing Strategies

I conclude this lecture with a few brief remarks concerning solar observing strategies at the VLA. Of course, the details of one's solar observing program depend very much on the objective of the experiment. I nevertheless mention several points to keep in mind when formulating an observing program.

4.1. Imaging the quiet Sun.

4.1.1. Choice of array configuration.

Experience has shown that the largest array configurations are of limited use for imaging the quiet Sun. First, aliasing due to undersampling relevant parts of the radio brightness distribution becomes acute. The net result is that confusion is even more severe in the largest array configurations and the resulting images more difficult to interpret.

Second, the Sun is near enough to the Earth that one must be aware that it is in the near field for certain array configurations and frequencies. Recall that the far-field approximation allows us to treat the incoming radiation as a plane wave—when this assumption is violated, the error in the phase is given roughly by $2\pi\nu b^2/Rc$. Here, ν is the frequency, R is the distance to the Sun, c is the speed of light, and b is the projected baseline. If the Sun were a point source, this would cause no particular difficulty since the error would be a simple, antenna-based phase shift that could be removed by self-calibration. But as the Sun is an extended source, the error cannot be corrected so simply. For the purposes of estimating when near-field effects are important, I assume that we can safely ignore them when the phase error given by the above expression is less than that introduced by phase-calibration transfer effects, i.e., so long as the effect is less than ~ 0.1 radians.

Figure 21–9 displays the result, demonstrating that while one need not worry about near-field effects in the **C** and **D** configurations at any frequency now supported by the VLA, they become increasingly important at the higher frequencies and for the longest baselines in the **B** configuration. In the **A** configuration, all frequencies above 327 MHz (90 cm) are affected to some degree.

4.1.2. Multiple frequencies.

Some quiet-Sun imaging experiments require the use of two or more frequency bands. While one can make use of subarrays (see below), this is rarely desirable for quiet-Sun imaging because of the price one pays in reduced u-v coverage. It is usually wiser to share time between the frequencies of interest, cycling them frequently enough to sample the longest baselines adequately.

The only mode of VLA operation that supports the use of two frequency bands simultaneously requires that one of the bands be 90 cm. The reason that this is possible is that the 90-cm feed is a crossed dipole located at the prime focus. Hence one could, for example, observe at 20 and 90 cm simultaneously.

4.1.3. Bandwidth synthesis.

An observational strategy that should be used more in future is bandwidth synthesis. It will be useful both to those interested in overcoming the problems associated with source variability (by imaging on short timescales) and to those interested in tracing the evolution of time-variable phenomena. It requires that one sample several frequencies within a given frequency band to increase the effective u-v coverage.

Figure 21–9. The heavy line demarcates those baselines and frequencies for which near-field effects become significant. The diagonal dashed lines indicate the various array configurations at the VLA.

4.2. Microwave bursts.

4.2.1. Snapshot imaging.

Because one is usually interested in imaging a microwave burst source on the shortest possible timescales, one must rely on the instantaneous u-v coverage of the VLA to image the source. As described in Section 3.5, snapshots are adequate to image microwave burst sources in most cases. Snapshots fail when the source is too large or complex. While it is true that phase calibration errors are largest when $T_{\rm ant}$ is elevated far above the quiet-Sun value, self-calibration techniques are applicable since the phase error is antenna-based.

4.2.2. Choice of array configuration.

The considerations which discouraged the use of the VLA **A** and **B** configurations for quiet-Sun imaging do not apply to the same extent for microwave burst sources, yet some care must be exercised in selecting the best array configuration. As stressed above, near-field effects are most severe in the large array configurations, yet both the **A** and **B** configurations may be profitably used at 20 and 90 cm.

4.2.3. The use of subarrays.

It is often desirable to image the source at two or more frequencies simultaneously. For example, the spectral peak of impulsive microwave bursts generally

Figure 21–10. An example of two-frequency imaging of a solar microwave burst during rise, peak, and decay phases of the event. The top row of panels represents total intensity contours of the 4.9 and 15 GHz sources, while the second and third rows represent the degree of circular polarization at each phase for 15 GHz and 4.9 GHz, respectively. [From Dulk, Bastian, and Kane (1986). Reprinted with permission from the *Astrophysical Journal.*]

lies between 3 and 10 GHz (Švestka 1976). It is frequently the case, therefore, that the burst source is optically thick at 6 cm and optically thin at 2 cm. To image a burst source at more than one frequency requires parcelling the array into two or more subsets of antennas. So long as reasonable care is taken in selecting with which subarray an antenna is affiliated, so that the u-v coverage provided by each is adequate, it is usually possible to image microwave burst sources at multiple frequencies fairly reliably. Figure 21–10 shows an example.

Acknowledgments

I thank G. Dulk and D. Gary for valuable comments and D. Bagri, P. Crane, R. Gonzalez, P. Lilie, and K. Sowinski for useful discussions.

References

Archer, J. W. (1979), "Modifications to the VLA front ends for solar observations", Addendum to VLA Electronics Memorandum No. 181, NRAO.

Bregman, J. D. and Felli, M. (1976), "The use of the Westerbork Synthesis Radio Telescope for solar observations at 21 cm", *Astron. Astrophys.*, **46**, 41–48.

Christiansen, W. N. and Warburton, J. A. (1955), "The distribution of radio brightness over the solar disk at a wavelength of 21 centimetres. III. The quiet Sun—two-dimensional observations", *Aust. J. Phys.*, **8**, 474–486.

Croom, D. L. (1971), "Solar microwave bursts as indicators of the occurrence of solar proton emission", *Solar Phys.*, **19**, 152–170.

D'Addario, L. R. (1979), "Observing the Sun with the VLA", VLA Scientific Memorandum No. 130, NRAO.

Dulk, G. A., Bastian, T. S., and Kane, S. R. (1986), "Two-frequency imaging of microwave impulsive flares near the solar limb", *Astrophys. J.*, **300**, 438–447.

Gary, D. E. and Hurford, G. J. (1987), "Multifrequency observations of a solar active region during a partial eclipse", *Astrophys. J.*, **317**, 522–533.

Gary, D. E. and Zirin, H. (1988), "Microwave structure of the quiet Sun", *Astrophys. J.*, **329**, 991–1001.

Habbal, S. R. and Harvey, K. L. (1988), "Simultaneous observations of 20 centimeter bright points and He I 10830 Å dark points in the quiet Sun", *Astrophys. J.*, **326**, 988–996.

Kattenberg, A. and Palagi, F. (1983), "One-dimensional high time resolution solar observations with the Westerbork Synthesis Radio Telescope", *Astron. Astrophys.*, **125**, 1–5.

Kundu, M. R. (1965), *Solar Radio Astronomy*, John Wiley & Sons, New York.

Kundu, M. R., Rao, A. P., Erskine, F. T., and Bregman, J. D. (1979), "High-resolution observations of the quiet Sun at 6 centimeters using the Westerbork Synthesis Radio Telescope", *Astrophys. J.*, **234**, 1122–1136.

Perley, R. A. and Erickson, W. C. (1984), "A proposal for a large, low frequency array located at the VLA site", VLA Scientific Memorandum No. 146, NRAO.

Švestka, Z. (1976), *Solar Flares*, D. Reidel (Dordrecht, Holland).

22. SYNTHESIS IMAGING OF SPATIALLY COHERENT OBJECTS

K. R. ANANTHARAMAIAH†,
T. J. CORNWELL, AND RAMESH NARAYAN‡

An important assumption that enters into the theory of synthesis imaging (see Lecture 1 by Barry Clark) is that the source being observed is spatially incoherent. Although there are some situations in which radio sources exhibit spatial coherence, synthesis imaging of spatially coherent objects has not been fully explored. Here we discuss some simple concepts associated with the imaging of such objects, and we describe a recent experiment that we have carried out with the VLA. A more detailed discussion can be found in Cornwell, Anantharamaiah and Narayan (1988a, b).

1. SPATIALLY INCOHERENT AND SPATIALLY COHERENT SOURCES

Let us consider an extended source. This source is said to be *spatially incoherent* if the radiation fields emitted by different parts of the source are independent of one another, i.e., if their phases and amplitudes are both uncorrelated. When the radiation from different parts of such a source arrives at the location of the observer, the component fields add incoherently. In other words, the intensities add (but the electric field amplitudes do not). Almost all extended radio sources in the sky belong to this category, and all the theoretical and practical aspects of imaging discussed in the previous lectures apply to such objects. The radiation emitted by an incoherent source is completely described by giving its intensity $I(P)$, which is a real positive number, as a function of position P in the source.

An extended source is considered *spatially coherent* if there are phase and amplitude relationships among the radiation fields received from different parts of the source. Such sources are not common. When the radiation contributions from two parts of such a source arrive at the location of the observer, they will add coherently and interfere; the voltages or fields, not the intensities, add together.

To describe the radiation emitted by a spatially coherent source, we need to consider a more general quantity than the intensity $I(P)$. We need the concept of the *mutual intensity* (e.g., Goodman 1985),

$$J(P_1, P_2) = \left\langle E(P_1)E^*(P_2)\right\rangle \ , \tag{22–1}$$

where P_1 and P_2 are two points in the source, $E(P)$ is the instantaneous complex radiation field at the point P and $\langle \cdots \rangle$ represents an expectation value, i.e., an

† On leave from Raman Research Institute, Bangalore 560080, India

‡ Steward Observatory, University of Arizona, Tucson, AZ 85721; supported in part by NSF grant AST–8814725.

average over many realizations of the wave field at the source. We notice that $J(P_1, P_2)$ differs from $I(P)$ in two ways—it is a function of two positions instead of just one, and it is in general a complex number for each pair of positions. (Generally speaking, we must also allow for a time lag τ between $E(P_1)$ and $E(P_2)$ and consider the *mutual coherence* $J(P_1, P_2, \tau)$, but we can avoid this complication under the so-called *quasi-monochromatic* condition, which amounts to observing with a sufficiently small bandwidth.) Even for a coherent source, we can continue to define the intensity $I(P) \equiv J(P, P)$, and this is still a real positive quantity; the difference is that the intensity is no longer a complete description of the radiation emitted by the source.

It is useful to normalize the mutual intensity and to define the *complex coherence factor*

$$\mu(P_1, P_2) = J(P_1, P_2) / \sqrt{I(P_1) I(P_2)} . \tag{22–2}$$

(The more general version of this quantity, including a time lag, is called the *complex degree of coherence*, $\gamma(P_1, P_2, \tau)$.) We may now define two types of spatial coherence. We call a source *perfectly coherent* if $|\mu| = 1$ for all pairs of points in the source. More generally, when $|\mu| < 1$ for some pairs, we call the source *partially coherent*. An incoherent source is, of course, described by the condition that $\mu(P_1, P_2) = 0$ for all pairs of points with $P_1 \neq P_2$.

We now note a major simplification for a perfectly coherent source. The distribution of radiation emitted by such a source can be completely described by a set of complex numbers $E'(P)$, each a function of only one position P (instead of two as for the mutual intensity). Indeed, $E'(P)$ represents the complex radiation field at P, and the prime is used to indicate that the phase of the field is measured with respect to the phase at some chosen point in the source (this reference point may be chosen arbitrarily). Because of this simplification, the mutual intensity of a perfectly coherent source can be factored into the form

$$J(P_1, P_2) = E'(P_1) E'^*(P_2) . \tag{22–3}$$

Furthermore, if the quasi-monochromatic condition is satisfied, this factorization property is maintained even during the propagation of radiation from a perfectly coherent source to the observer, and therefore Equation 22–3 will be valid even at the telescope, where P_1 and P_2 now refer to the locations of the two elements of an interferometer.

We shall later give an example of an astronomical source that is perfectly coherent. In the laboratory, a Young's double slit, illuminated by a point light source, is a good example of a perfectly coherent source. If the light source is extended, then the slits behave like a partially coherent source.

2. Synthesis Imaging of a Spatially Incoherent Source

This has been discussed extensively in the previous lectures. A two-element interferometer measures the amplitude and phase of one spatial frequency (one point in the *u*-*v* plane) of the brightness distribution of the source. Earth rotation sweeps the interferometer through a range of spatial frequencies (a track in the *u-v* plane). An array of interferometers (351 interferometers for the VLA), together with Earth rotation, measures visibilities at many spatial frequencies that can then be Fourier transformed to estimate the brightness distribution of the source (to yield a "dirty" image). Deconvolution and self-calibration techniques can then be used to correct and improve this image.

For this discussion, the important point to note is that for a spatially incoherent source, a measured visibility (the complex amplitude and phase at some spatial frequency) is a function only of the projected interferometer spacing and its orientation. It is independent of the location of the two antennas. In other words the image of a spatially incoherent object (in practice, any extended object in the sky) that we obtain with the VLA will be unaffected even if we 'bodily' shift the VLA to another location (keeping the orientation the same).

3. Synthesis Imaging of a Spatially Coherent Source

To appreciate the differences that arise in imaging a spatially coherent object, consider a two-element interferometer operating in the radiation field of a perfectly coherent double source (see Fig. 22–1). This source is similar to a Young's double-slit in the laboratory. The radiation from the two sources A and B is perfectly coherent in the sense that there is a fixed phase relationship. As a result, the radiation field on the observer plane (where the interferometer is located) has a sinusoidal fringe pattern. Let the two elements of the interferometer be located at C1 and C2 with a separation d. As the fringe pattern has finite amplitude at C1 and C2, the interferometer will measure a finite value for the visibility amplitude. Now let us bodily shift this interferometer to location D1 and D2. Since we have not changed the orientation or the distance between the antennas, the interferometer will again measure the same "*u-v* point". However, now D1 is located at the null of the fringe pattern, so the interferometer measures zero for the visibility amplitude. This results from the spatially coherent nature of this simple source. If A and B were now to be replaced by an incoherent double source, there would be no fringe pattern and the interferometers C1–C2 and D1–D2 would measure the same visibility (except for the noise term).

Therefore, for a spatially coherent source the visibility measured by an interferometer depends not only on the baseline (the separation and orientation) but also on the location of the antennas. In other words, while the imaging of a spatially incoherent source with an interferometer is a two dimensional problem (the coherence function depends only on the baseline or position in the *u-v* plane), the imaging of a spatially coherent source is four dimensional (the coherence function depends on the positions (x_1, y_1) and (x_2, y_2) of the antennas). This is true for perfectly coherent and partially coherent sources.

If we consider a spatially coherent source that is more complex, then the fringe pattern in the two dimensional observer plane will be more complicated than the simple sinusoidal pattern discussed above. There will be many maxima and minima, and perhaps some nulls in the pattern. A synthesis array located on the observer plane will sample this complex fringe pattern at the locations of the antennas. Normally, in a synthesis array of N elements, all the $N(N-1)/2$ possible correlations between the antennas are recorded. These supply indepen-

Figure 22–1. A two-element interferometer observing a perfectly coherent double source at two locations on the observer plane (C1–C2 and D1–D2).

dent information (assuming the array is non-redundant) for an incoherent source. However, for a perfectly coherent source, the factorization property described by Equation 22–3 makes the $N(N-1)/2$ correlations interdependent. They can be factored into N independent terms, each complex-valued, that represent the amplitude and phase of the fringe pattern at the locations of N antennas (we may use one antenna as a phase reference).

In certain situations, the complex field distribution of the spatially coherent source is directly obtainable by Fourier transformation of the complex amplitudes at the locations of the antennas. If N is large enough and if the fringe pattern in the observer plane is sufficiently well sampled, this Fourier transform may be carried out to obtain the true field distribution of the coherent source. However, in synthesis arrays like the VLA, N is small (only 27(!)) and therefore it would be difficult to obtain the correct image of a spatially coherent source. We can envisage deconvolution algorithms involving "dirty images" made from the complex amplitudes at the antennas and "dirty beams" made from the N locations (in the the x-y plane) of the antennas. But, given the small numbers, the results may not be good.

Even though using the VLA we cannot expect good images of a spatially coherent source, we can use the concepts discussed above to test for the presence of perfect coherence of the radiation field. The test is to observe a perfectly coherent source and to see whether the $N(N-1)/2$ complex correlations that we measure can be factored into just N independent complex numbers. As described above, this happens only if the source is perfectly coherent. We later describe a practical way to make this test using the conventional self-calibration procedure.

4. Spatially Coherent Astronomical Sources

An unresolved point source can be considered to be perfectly coherent. However, there is then no fringe pattern on the observer plane, as the radiation arrives from only one direction. All interferometers will measure the same visibility and, for the present discussion, this is not an interesting case. (For an unresolved source it is well known that the $N(N-1)/2$ correlations can be factored into N independent complex-valued terms that represent the complex gains of the N elements of the synthesis array. This is the basis on which the antenna gains are calibrated.)

It is hard to find extended sources that are intrinsically spatially coherent. Radiation must originate in different parts of the source through some mutually cooperative phenomenon. Maser sources are coherent sources. However, by their very nature, they are spatially coherent along the line-of-sight and therefore, most maser sources are intrinsically point-like. To date, no extended source is known that is spatially coherent because of an internal physical process.

In radar reflection experiments (e.g., Grossman and Muhleman 1988), the radiation reflected by the planets and the Moon may be spatially coherent, under certain conditions.

5. Spatial Coherence by Radio Wave Scattering

It was mentioned earlier that a Young's slit arrangement corresponds to a spatially coherent source. In the laboratory, one can generalize this set-up by having multiple holes (instead of merely two slits) being illuminated by a distant point source. By varying the sizes and locations of the holes and by adding different phase shifts at the holes (using different thicknesses of a refracting medium), it is possible to generate an arbitrarily complex coherent source in the laboratory. Something closely analogous to this occurs in radio astronomy when the radiation from a distant radio source is scattered by random density fluctuations in an ionized medium *en route* to the observer (see Fig. 22–2).

Figure 22–2. An extended spatially coherent source generated by radio wave scattering by an ionized screen. In the absence of the scattering screen, the source appears like a point source at the observer plane. The scattering generates varying fringes on the observer plane, and the source appears extended when observed with an imaging instrument. For short integration times the broadened source will exhibit spatial coherence.

Radio scattering by ionized matter is seen in three media in astronomy—in the interstellar medium of the Galaxy, in the interplanetary medium of the solar system, and in the ionosphere of the Earth. The first two are of particular interest to us since the conditions often correspond to so-called *strong scattering.* In this regime of scattering, each point on the observer plane receives radiation from many independent scattering blobs (multi-path propagation), each contribution having an independent phase shift added to it by the associated blob.

The situation is therefore similar to the multiple-hole laboratory experiment mentioned above. If the characteristic scattering angle of the scatterers is θ, then radiation is received at the telescope over a range of angles approximately equal to some fixed angle θ and one says that the source is *scatter-broadened* by this angle. If the background source is compact enough (we discuss this requirement in more detail in Section 7), and if the bandwidth of the observations is small enough (the quasi-monochromatic condition), then the radiation received from different scattering blobs will be perfectly coherent and there will be deep intensity modulation on the observer screen. If now there is relative motion between the scatterers and the line connecting the observer and the source, one obtains strong temporal fluctuations in the intensity, called *scintillation.* In the three astronomical media (mentioned above) where radio scintillation occurs, the phenomenon is called interstellar scintillation (ISS) (e.g., Rickett 1977), interplanetary scintillation (IPS) (e.g., Readhead, Kemp and Hewish 1978) and ionospheric scintillation, respectively.

When strong scintillation occurs, the radiation emanating from the various scattering blobs can be viewed as being emitted by a perfectly coherent source. For IPS at small solar elongations, the scatter-broadened disk is large enough to be resolved by the VLA. This gives us the opportunity to observe a scatter-broadened image under conditions when it behaves like a perfectly coherent source. By the discussion in Section 1, the radiation at the radio telescope will then be perfectly coherent and can be factored in the manner shown by Equation 22–3. We describe next a test of this property that we have carried out using the VLA. Here we note that such a test needs to be done in a time short compared to the time it takes an interference fringe to sweep past the telescope (the scintillation time-scale). For IPS, the solar wind moves at a velocity of 200 to 400 km s^{-1} (Ekers and Little 1971), and the fringe pattern moves over the extent of the VLA (35 km in the **A** configuration) in about 100 milliseconds. If the imaging is done with integration times of a few tens of milliseconds, then it will correspond to spatially coherent conditions. For longer integrations, the visibilities get averaged over many fringes, and the scatter-broadened source will behave like any spatially incoherent extended source.

6. A VLA Experiment

As noted above, to observe spatial coherence in a radio source that has been scatter-broadened by the solar wind, the integration time should be a few tens of milliseconds. Normally, the VLA data recording system allows a minimum integration time of 6.66 seconds if all the correlations between the antennas are recorded. There is, however, a special technique that allows us to obtain high time resolution images with the VLA.

6.1. High time resolution "snapshot images" with the VLA.

Although the data dump from the correlator can occur only once every 6.66 s, the correlator can integrate for times as short as 92 μs when gated with a train of electrical pulses. When the pulse is "on", the correlator performs the multiplication, and when the pulse is "off", it ignores the input signals.

Figure 22–3. A 5 msec snapshot image of the source NRAO 150 made using the correlator gating technique. The observing frequency was 4.885 GHz. Contour levels are −4, −2, 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, and 90 percent of the peak brightness, 9.747 Jy/beam. The r.m.s. noise in the image is 97 mJy/beam.

We therefore generated a train of pulses, each of width 10–40 ms and with a period of 6.66 s. By gating the correlator with these pulses snapshot images with integrations of 10 to 40 ms could be obtained once every 6.66 seconds.

The VLA receiver system has two data paths that can be tuned independently to any frequency within an observing band. These are referred to as the AC and BD channels. The correlators for the two sets of channels can be gated independently. In the observations described below, we tuned both the channels to the same frequency and applied the gating train of pulses only to

Figure 22–4. Time-averaged images of 3C 279 made at λ20cm when the source was at different angular distances from the Sun. The elongation is shown for each frame. The peak flux densities in the images are, from left to right, 1.55, 3.34, 5.44, and 6.35 Jy/beam, respectively. The contour levels are -10, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 percent of the peak. The beam size was $\sim 1\farcs5 \times 1\farcs4$

the AC channels. Therefore, the BD channels provided time-averaged images, whereas the AC channels provided the required snapshots. In observations at short wavelengths (e.g., λ2cm and λ6cm), it is necessary to correct for atmospheric phase variations across the array. With the above scheme this was done by self-calibrating the BD channels in phase only, and applying the resulting corrections to both the AC and BD channels.

To show that the VLA can image successfully in short integration times, we observed the strong source NRAO 150 using the above technique when the source was far away from the Sun (elongation $> 40°$). The observations were made in the **A** configuration. A gating pulse width of 5 ms was used. The observing frequency was 4.885 GHz, and the bandwidth was 50 MHz. Figure 22–3 shows the final image after phase self-calibration and deconvolution (using 'CLEAN'). The root-mean-square noise in this image is 97 mJy/beam, giving a signal-to-noise ratio of about 100 on this 9.7 Jy source.

6.2. Observations of 3C 279.

The apparent position of the Sun nears that of the strong radio source 3C 279 each October. During these occasions, the radiation from this source is subject to strong scattering by density fluctuations in the solar wind plasma. 3C 279 is almost point-like, but with low-brightness extended emission that can be seen only in high dynamic range images. For our purposes it may be considered to be a point source. The angular scatter-broadening close to the Sun is significant—a few arcseconds at centimeter wavelengths (Armstrong *et al.* 1986). In October 1987, the VLA was in the **A** configuration and therefore provided the necessary angular resolution at centimeter wavelengths to resolve the scatter-broadened image of 3C 279.

We observed this source on eight occasions at wavelengths λ2cm, λ6cm, λ20cm, and λ90cm. Gating pulse widths of 10 ms and 40 ms were used. The time-averaged images (from the BD channels) obtained at λ20cm, λ6cm, and λ2cm at different elongations are shown in Figures 22–4, 22–5 and 22–6. The

Figure 22–5. Time-averaged images of 3C 279 made at λ6cm when the source was at three different angular distances from the Sun as shown. The peak flux densities in the images are, from left to right, 0.91, 7.2, and 8.5 Jy/beam respectively. In units of percentage of the peak, the contour levels for the first image are 10 to 100 in steps of 10; for the second and third images they are, 10, 30, 50, 70, and 100. The beam size was $\sim 0\farcs43 \times 0\farcs38$.

averaging time for these images is ~ 10 min. These images show the increase in scattering strength and the resulting increase in the angular broadening as the line of sight passes closer and closer to the Sun's disk. The magnitude of the angular broadening increases with wavelength, as λ^2. Note also that the angular broadening is asymmetric, indicating an anisotropy in the scattering medium. These results are consistent with earlier observations by Armstrong *et al.* (1986).

6.3. A test for spatial coherence in the snapshot images.

The snapshot images of 3C 279 obtained above are expected to show perfect spatial coherence if three conditions are satisfied:

(1) The source must be sufficiently point-like to be unresolved by an aperture the size of the scattering disk. 3C 279 is compact enough to satisfy this requirement.

(2) The integration time must be short compared to the scintillation time scale.

(3) The bandwidth used must satisfy the quasi-monochromatic condition. In the present application this corresponds to using a bandwidth that is small compared to the decorrelation bandwidth of the scintillations. (This bandwidth is given by the reciprocal of the difference between the propagation times of radio waves from the extremes and center of the scatter-broadened disk.)

Figure 22–6. Time-averaged images of 3C 279 at λ2cm at two elongations observed on Oct. 7 and Oct. 10. The peak flux densities in the two images are 9.7 and 16.0 Jy/beam, respectively. The contour levels are, in units of percent of the peak, 10 to 100 in steps of 10.

These conditions were satisfied for our observation at λ2cm on October 7th, when the elongation of 3C 279 was $0^{\circ}\!.9$ and the bandwidth used was 50 MHz. In the first row of Figure 22–7 we show three snapshot images made on this day with an integration time of 40 ms.

These images were made using the standard procedures that are appropriate for spatially incoherent sources, i.e., the 351 measured correlations were Fourier transformed to get a "dirty image" that was then 'CLEAN'ed using the corresponding "dirty beam". At this point, the meaning of these images is unclear since, as discussed in Section 3, the image of a perfectly coherent object should be obtained by a Fourier transform of the complex amplitudes at the location of the antennas and not by a Fourier transform of the correlations between the antennas. Nevertheless, we have used the standard procedures and we note that some of the fine structure and the spottiness in these images is a result of this. We can perhaps say that the appearance of the fine structure in these images is a sign of the presence of spatial coherence in the object (assuming, of course, that there were no instrumental problems). In the absence of spatial coherence these images would have been smooth and would have had the shape of the scatter-broadened disk ("seeing disk").

As discussed in Section 3, the real test for the presence of perfect coherence is to show that the 351 correlations that we measure can be factored into just 27 independent complex numbers that can be regarded as the complex amplitudes of the fringe pattern at the location of the antennas. These 27 complex numbers

Figure 22–7. Examples of snapshot images of 3C 279 made at λ2cm on Oct. 7, when the elongation was $0\overset{\circ}{.}9$. *(First row)* 40 msec snapshots from AC channels (see text). *(Second row)* Amplitude and phase self-calibrated snapshot images. *(Third row)* Long-integration (6.66 sec) images from BD channels. *(Fourth row)* Amplitude and phase self-calibrated long-integration images. The contour levels are −90, −80, −70, −60, −50, −40, −30, −20, −10, −8, −6, −4, −2, 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, and 90 percent of the peak brightness.

can also be thought of as the complex gains of the antennas. These two cases are identical and cannot be distinguished. Therefore, if we now apply an amplitude and phase self-calibration with a point source model to these data, then we ought to obtain 27 complex gain “corrections” for the antennas. These should represent the complex amplitudes of the fringe pattern. If we apply these “corrections” and again construct the image (using standard procedures), then we expect to get a point-source. This is a test for the presence of perfect coherence. We followed this procedure for the three images and the resulting “corrected” images are shown in the second row of Figure 22–7. These images show just a point source, showing that the snapshot images were indeed perfectly coherent.

If the above procedure is applied to a spatially incoherent source, then we would not expect to get a point source in the corrected image. Images of 3C 279 made using the BD channels (where no gating was applied) with an integration of 6.66 seconds are not expected to show spatial coherence. They will be spatially incoherent because the fringes get averaged. Three such images are shown in the third row of Figure 22–7. These images do not have any fine structure (unlike the first row in Fig. 22–7) and already show that there is no spatial coherence. The corrected images after self-calibration are shown in the fourth row. As expected, they are not point sources.

In the left column of Figure 22–8, we show the correction factors (or the complex gains) obtained from self-calibration of the three snapshots of Figure 22–7 (first row). The complex gain for each antenna is plotted at its position as projected on a plane perpendicular to the direction of the source. These gains have the expected behavior; they show fringes running across the array. The complex gains estimated for the long-integration images (Fig. 22–7, third row) are shown in the right-hand column of Figure 22–8. No fringes are visible—phases are zero and the drop in amplitudes for the extreme antennas is a result of using a point source model for an extended source during the self-calibration.

Thus a test for perfect coherence in an extended source is to show that self-calibration (amplitude and phase) with a point source model results in a point source. In other words, in the presence of perfect spatial coherence in an extended source, the self-calibration procedure always yields a point source. Therefore, one must be careful about using self-calibration in any search for coherent emission in radio sources.

7. Analogy with an Aberrated Single Dish

The generation of coherence through scattering by plasma irregularities is analogous to a similar phenomenon, namely, when the scattering occurs in a single-dish radio telescope with significant aberrations. The irregularities in the dish surface scatter incident radio waves and produce a fringe pattern, similar to the one discussed earlier, at the focal plane. Cornwell and Napier (1986, 1988) have explored the attributes of this scattered radiation field. They showed that the coherence function in the focal plane of a radio telescope (which is similar to the correlations measured by different interferometers on the observer plane in our experiment) can be used to obtain information about both the scattering surface and the intrinsic source structure. This raises the interesting

Figure 22–8. Complex gains of antennas obtained from amplitude and phase self-calibration. *(Left column)* Complex gains for snapshot images in the first row of Figure 22–7. *(Right column)* complex gains for long integration images in the third row of Figure 22–7. The length of each vector denotes the amplitude, and the position angle denotes the phase. A unit-amplitude, zero-phase vector is shown in the bottom left corner of each plot.

possibility that something similar may be possible even in the scintillation case that we have studied. If this turns out to be possible, then the prospects are exciting as the effective "dish" in the scintillation case is the scattering disk in the scattering medium, whose diameter can be much larger than the aperture of the telescope. (This is particularly true for ISS, where the scattering disk is often more than 1 a.u. across.) Consequently, by the careful analysis of snapshot images, it may be possible to achieve resolutions far exceeding that appropriate to the telescope. Note that the surface of the dish which causes the scattering in the Cornwell–Napier analysis is stationary. Therefore, for their analysis to apply in our problem, we must effectively freeze the scattering screen. This is, of course, what we do when we observe with short integration times of 10–40 ms.

The test for perfect coherence, described in the previous section, requires that the radiation incident on the scattering screen be perfectly coherent. In other words, that the source be unresolved over baselines comparable to the size $\sim \theta L$ of the scattering disk, where θ is the scattering angle and L is the distance between the scattering screen and the observer. The interesting case, which is appropriate for obtaining high angular resolution as suggested above, occurs when the source is partially resolved over these baselines. The radiation incident on the screen is then partially coherent and the coherence function contains information about the structure of the source. The analyses of Cornwell and Napier (1986, 1988) show that it is possible to obtain this coherence function, and therefore the structure of the source, by making measurements on the observer plane.

The following argument explains how it is possible to obtain a resolution corresponding to the size of the scattering disk even though the coherences are measured only over the size of the synthesis array. Consider two scattering blobs in the scattering screen separated by a distance D that is much larger than the size d of the synthesis array. Let radiation be incident on these blobs from a distant source. If the source is unresolved on a baseline of length D, then the radiation arriving at the two blobs will be perfectly coherent. On scattering by the blobs, the radiation will arrive at the observer plane retaining its perfect coherence for narrow observing bandwidths (the quasi-monochromatic condition.) The fringes that are created by the perfect coherence can be detected by measuring the coherence function in the observer plane as described earlier. In fact, the fringes could be detected even with a single antenna if the blobs move in a lateral direction (e.g., outward flow of the solar wind), since the movement will cause a fluctuation in the intensity at the location of the antenna and lead to scintillation. This is the basis of the early IPS observations with single antennas which measured the angular diameters of compact radio sources. If the source is resolved over a baseline of length D, then the radiation arriving at the two blobs is incoherent and there will be no fringes on the observed plane. This simple experiment shows that just the detection or non-detection of the fringes can tell whether the source has structure on scales $\sim \lambda/D$, where λ is the observing wavelength. One can therefore visualize that, given a collection of scattering blobs with a range of separations, it should be possible to obtain the intrinsic *structure* of the source by measuring the full coherence function of the resulting fringe pattern. Further work is needed in this regard.

The resolution attainable by this method using solar-wind scattering corresponds to baselines of at most a few thousand kilometers. This baseline length is easily obtained with present-day VLBI techniques. However, in interstellar scattering the baselines could range up to a few astronomical units. Wolszczan and Cordes (1987) were the first to use such a technique with a single antenna to resolve the magnetosphere of a pulsar. They used scattering blobs in the interstellar medium in their experiment and aptly called their technique "interstellar interferometry".

Acknowledgments

We thank K. Sowinski for suggesting the correlator-gating technique used in the VLA observations, and for advice about the VLA on-line computer system.

References

Armstrong, J. W., Coles, M., Kojima, M., and Rickett, B. J. (1986), "Solar wind observations near the Sun", in *The Sun and the Heliosphere in Three Dimensions*, R. G. Marsden, Ed., D. Reidel (Dordrecht, Holland), pp. 59–64.

Cornwell, T. J., Anantharamaiah, K. R., and Narayan, R. (1988a), "Spatial coherence in a scatter-broadened image—observations of 3C 279 close to the Sun", in *Radio Wave Scattering in the Interstellar Medium*, J. M. Cordes, B. J. Rickett and D. C. Backer, Eds., AIP Conference Proceedings **174**, American Institute of Physics, New York, pp. 92–96.

Cornwell, T. J., Anantharamaiah, K. R., and Narayan, R. (1988b), "The propagation of coherence in scattering—an experiment using inter-planetary scintillation", *J. Opt. Soc. Amer.*, submitted.

Cornwell, T. J. and Napier, P. J. (1986), "Self-correction of telescope surface errors using a correlating focal plane array", in *Radio Astronomy from Space*, K. W. Weiler, Ed., Workshop Vol. No. 18, NRAO (Green Bank, WV), pp. 215–219.

Cornwell, T. J. and Napier, P. J. (1988), "The focal plane coherence function of an imaging antenna and its use in measuring and correcting aberrations", *Radio Science*, **23**, 739–748.

Ekers, R. D. and Little, L. T. (1971), "The motion of the solar wind close to the Sun", *Astron. Astrophys.*, **10**, 310–316.

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

Grossman, A. and Muhleman, D. O. (1988), in preparation; reported by J. Eberhart, *Science News*, **134** (1988) p. 170.

Readhead, A. C. S., Kemp, M. C., and Hewish, A. (1978), "The spectrum of small-scale density fluctuations in the solar wind", *Mon. Not. Roy. Astron. Soc.*, **185**, 207–225.

Rickett, B. J. (1977), "Interstellar scattering and scintillation of radio waves", *Ann. Rev. Astron. Astrophys.*, **15**, 479–504.

Wolszczan, A. and Cordes J. M. (1987), "Interstellar interferometry of the pulsar PSR 1237+25" *Astrophys. J.*, **320**, L35–L39.

23. NOISE IN IMAGES OF VERY BRIGHT SOURCES

K. R. Anantharamaiah[†],
Ronald D. Ekers[‡], V. Radhakrishnan[§],
T. J. Cornwell and W. Miller Goss

This lecture discusses the characteristics of noise in synthesis images of very bright sources. For most radio sources and all existing synthesis arrays, the antenna temperature T_A due to the source in an individual element of the array is usually much smaller than the receiver temperature T_R. The fluctuations at the output of an antenna are therefore usually dominated by the receiver noise. Much of the existing analysis of the distribution of noise in a synthesized image is, therefore, concerned with the situation in which the source can be considered as "weak" (e.g., Christiansen and Högbom 1985; Högbom and Brouw 1974; Rogers 1976; Thompson *et al.* 1986; Crane and Napier, Lecture 7). However, for a few sources such as Cassiopeia A, Cygnus A and the Crab Nebula, the antenna temperature, particularly at wavelengths longer than 20 cm, can be much higher than the receiver temperature in an individual element. For example, at $\lambda =$ 20 cm, the antenna temperatures due to these sources at the VLA antennas are about 240 K, 155 K, and 100 K, respectively, whereas the receiver temperature $T_R \approx 65$ K. Such sources are considered "very bright" because the fluctuations at the output of an antenna are dominated by the source noise rather than the receiver noise. Furthermore, when we consider the total collecting area of a synthesis array, the effective antenna temperature due to even moderately strong sources (a few tens of Jy) can exceed the receiver temperature. For the 27 VLA antennas, the effective antenna temperature due to a source is $\sim$ 2.7 K/Jy. Therefore, at $\lambda = 6$ cm, for example, the effective antenna temperature due to a 20 Jy source exceeds the nominal receiver temperature of 50 K. The "on-source" noise in a synthesized image can then differ from the "off-source" noise. This lecture discusses the distribution of noise in synthesis images of moderately bright and very bright sources.

In this lecture, we include in the receiver temperature all the additive sources of noise such as the spillover, waveguide losses, the atmospheric contribution and the 3 K sky background. We shall refer to the fluctuations due to the antenna temperature as *source noise*, and those due to the receiver temperature as *receiver noise*. In all discussions, we use units of temperature to represent noise.

[†] On leave from Raman Research Institute, Bangalore 560080, India.
[‡] Australia Telescope, CSIRO, P. O. Box 76, Epping, NSW 2121, Australia.
[§] Raman Research Institute, Bangalore 560080, India.

1. Imaging with a Single Dish

The standard radiometer equation for the r.m.s. noise at the output of a total power receiver connected to a single dish is

$$\Delta T = \frac{T_{\rm sys}}{\sqrt{B\tau}}\,, \tag{23–1}$$

where $T_{\rm sys} = T_A + T_R$, B is the bandwidth and τ is the integration time. The antenna temperature due to the source is given by

$$T_A = \frac{SA_e}{2k}\,, \tag{23–2}$$

where k is the Boltzmann constant, S is the flux density of the source and A_e is the effective collecting area of the antenna. $A_e = \eta A$, where η is the aperture efficiency and A is the physical area.

1.1. Weak source.

Here we have $T_A \ll T_R$, so $T_{\rm sys} \approx T_R$. The r.m.s. noise and the signal-to-noise ratio are given by

$$\Delta T = \frac{T_R}{\sqrt{B\tau}} \quad \text{and} \quad \frac{T_A}{\Delta T} = \frac{SA_e\sqrt{B\tau}}{2kT_R}\,. \tag{23–3}$$

Here, increasing the collecting area A_e of the antenna increases the signal-to-noise ratio.

1.2. Strong source.

In this case, $T_A \gg T_R$ and $T_{\rm sys} \approx T_A$. From Equations 23–1 and 23–2 the signal-to-noise ratio is

$$\frac{T_A}{\Delta T} = \sqrt{B\tau}\,, \tag{23–4}$$

which is independent of the collecting area A_e of the antenna. The meaning of this result is that, for very strong sources, the fluctuations at the output of the receiver are completely dominated by the source noise. The signal-to-noise ratio can be improved only by increasing the bandwidth or the integration time, or both.

Suppose we have a large antenna, with a consequently small beam, and we wish to observe a small region surrounding a strong source. We assume for simplicity that the sidelobe levels of the primary beam of the antenna are negligibly small. The required image can be made by taking scans over this region with suitable separation between the scans. It is easy to see that such an image will have noise proportional to T_R in regions where the beam does not intersect the source. However, on the source the noise will be much higher and will be proportional to $T_R + T_A \approx T_A$.

In this lecture, we want to explore the distribution of noise in an image of a similarly strong source made using a synthesis array. It is often said that we achieve the same result with a synthesis array as with a single aperture that is as large as the array. If this is strictly true, then, in the synthesis image of a strong source, we expect that the noise outside the source will be due to the receiver noise (T_R) and that the noise on the source will be due to the total noise ($T_A + T_R$). However, as we show here, the expected off-source noise in a synthesis image depends on how the image is obtained—(a) using a snapshot or (b) using sequential observations involving Earth rotation or movable elements. We show that in a snapshot it is possible, in principle, to achieve off-source noise that is only due to the receiver noise. In a long-integration synthesized image, however, the off-source will have the full contribution of the source noise as well as the receiver noise. We begin with a discussion of noise in a simple two-element interferometer.

2. Noise in a Two-Element Interferometer

A more complete discussion of this has been presented by Crane and Napier (Lecture 7). Here we summarize some of the results. For simplicity, we consider identical elements of effective collecting area A_e, with equal receiver temperatures (T_R) and a single multiplier. The general expression for the r.m.s. noise at the output of the multiplier has been derived by Crane and Napier. Converting their Equation 7–32 to units of temperature, we get

$$\Delta T = \frac{1}{\sqrt{B\tau}}\sqrt{T_A^2 + T_A T_R + \tfrac{T_R^2}{2}}. \qquad (23\text{–}5)$$

We now consider the signal-to-noise ratio for three separate cases.

2.1. Weak source.

Since $T_A \ll T_R$, the r.m.s. noise at the output of the correlator is given by

$$\Delta T \approx \frac{1}{\sqrt{2}}\,\frac{T_R}{\sqrt{B\tau}}\,. \qquad (23\text{–}6)$$

The correlated power is proportional to T_A. Comparing this to Equation 23–3, we notice that, although we have used two antennas each of collecting area A_e the reduction in noise is only $\sqrt{2}$. A single dish of collecting area $2A_e$ would have increased the signal-to-noise ratio by a factor of two. As pointed out by Crane and Napier, this loss in sensitivity for a correlation receiver occurs because the interferometer does not also use the autocorrelation of the signal from each antenna.

2.2. Strong unresolved source.

Here $T_A \gg T_R$ and we obtain

$$\Delta T \approx \frac{T_A}{\sqrt{B\tau}}\,, \qquad (23\text{–}7)$$

which is identical to Equation 23–4 and shows that the signal-to-noise ratio is independent of the collecting area. The correlated power will be proportional to T_A. By comparing Equations 23–6 and 23–7, Crane and Napier have noted an important difference between correlated and uncorrelated noise: correlated noise produces $\sqrt{2}$ times more fluctuations at the output of a correlator, compared to uncorrelated noise.

2.3. Strong, fully resolved source.

We consider a strong ($T_A \gg T_R$) extended source that is not resolved by the individual antennas but is resolved by the baseline that separates them. When a source is fully resolved by the interferometer the correlated power is zero. In other words, the noise fluctuations at the output of the two antennas are completely independent. This is the basis of intensity interferometers (e.g., Hanbury Brown 1974). Although the system temperatures of the two antennas are dominated by the contribution from the source, the noise fluctuations become independent if the antennas are separated by a distance that resolves the source. The noise contributions from the source behave as if they came from independent *receivers* in the two antennas and the r.m.s. noise at the output of the correlator has the same form as Equation 23–6, so that

$$\Delta T \approx \frac{1}{\sqrt{2}} \frac{T_A}{\sqrt{B\tau}} \,. \tag{23–8}$$

By comparing Equations 23–8 and 23–4, Crane and Napier note that the fluctuations at the output of an interferometer, in this case, are $\sqrt{2}$ lower than for a single antenna. Therefore, in the presence of a strong resolved source that dominates the system temperature, a synthesis array of N elements gives N times better signal-to-noise ratio for a weak source than a single antenna with a collecting area equal to that of the array. This is true only when the strong source is unresolved by the single large antenna and when it dominates the system temperature in the individual elements of the array. Such situations occur, for example, in solar observations where we can try to detect localized solar activity in the presence of very strong disk emission (see Lecture 21), or weak line emission in the presence of a strong extended continuum background. Another good example is the detection of the Crab pulsar in the presence of the strong extended continuum emission from the Crab Nebula.

3. Noise in Synthesis Images of Weak Sources

For the purpose of estimating noise in the off-source region of an image, a source can be considered weak when the antenna temperature due to the source in an individual element of the synthesis array is much smaller than the receiver temperature: $T_A \ll T_R$. This is true for most observations. An expression for the noise in the image for this case has been derived by Crane and Napier,

$$\Delta T = \frac{T_{\text{sys}}}{\sqrt{N(N-1)B\tau}} \,, \tag{23–9}$$

where N is the number of elements in the array, and we have considered a single IF. Intuitively, it is easy to understand the origin of this expression, by starting from Equation 23–6 which was derived by Crane and Napier for a single two-element interferometer. In a synthesis array of N elements there are $N(N-1)/2$ baselines. Since $T_A \ll T_R$, each baseline can be considered as making an independent measurement, because the receiver noise T_R is independent for each element. The noise in a final image that uses all possible correlations is therefore reduced by a factor $\sqrt{N(N-1)/2}$.

Equation 23–9 estimates the noise in the off-source region of the synthesis image of a weak source as defined above, i.e., $T_A \ll T_R$. A source can be considered moderately strong when the effective antenna temperature due to the *total collecting area* of the array exceeds the receiver temperature, i.e., $NT_A > T_R$. Note that a moderately strong source can still satisfy the weak source criterion, $T_A \ll T_R$. In these cases, the noise in the on-source region can be significantly different from that given by Equation 23–9. Section 4.3 discusses this further.

It is useful to clarify the meaning of r.m.s. noise in the context of an image. In a "dirty" image made by Fourier transforming the measured correlations, the r.m.s. fluctuations between different pixels in the image will be dominated by the sidelobe structure of the dirty beam. Equation 23–9 does not represent these fluctuations. It would, for example, represent the r.m.s. fluctuations in the value of a given pixel from one realization of the image to another.

4. Noise in Synthesis Images of Strong Sources

4.1. Strong unresolved source.

In this case, $T_A \gg T_R$. As seen in Section 2.2, for a strong unresolved source that dominates the system temperature, the r.m.s. fluctuations at the output of a two-element interferometer are given by (*cf.* Eq. 23–7)

$$\Delta T = \frac{T_A}{\sqrt{B\tau}}\,. \tag{23–10}$$

In a synthesis array of N elements, all the $N(N-1)/2$ interferometers will have identical fluctuations at the output. Therefore, unlike the weak source example discussed above, there is no reduction in noise when the different measurements are combined to form an image. Thus, we expect the noise in the final image to be given by Equation 23–10; the noise is independent of the number of elements in the array. No increase in the collecting area of the array, either by increasing the number of elements or by increasing the area of the individual elements, improves the signal-to-noise ratio. This is analogous to the case of a single dish with increasing area.

We now discuss the distribution of this noise in the image. By analogy with Equation 23–9, we expect Equation 23–10 to represent the noise at all points in the synthesis image of a strong unresolved source. If this is true, then it differs considerably from the distribution of noise in an image made using a single dish, as discussed in Section 1.2. In a single-dish image, this noise appears only at the position of the source. Is it possible to achieve the same result in a synthesis image? As we show below it is indeed possible to do so if the image is made in a 'snapshot' mode. However, in images made from long observations using either Earth rotation or movable elements, the noise at any point away from the source is given by Equation 23–10. This difference is a consequence of the changing pattern of the instantaneous "dirty beam" of a synthesis array, as a function of time. We consider three cases:

4.1.1. Filled array.

Consider first a filled array, which we can imagine to be a number of square dishes placed next to each other without any space between them. It is reasonable to assume that the sidelobe levels for this filled array are negligible. Let all the dishes be pointed at a single strong point source. There will then be noise in each receiver corresponding to the antenna temperature of the source in each of the telescopes. If we now 'phase' the array to point to the source, the combined output will be the same as we would have obtained from a single dish whose area equals that of the whole array, and pointed at the source. (This is equivalent to the phase center of an image obtained by recording all the correlations between the individual dishes and using a 'direct Fourier transform'.) On the other hand, if we phase the array to look at a point within the primary beam of each of the elements, but displaced from the source and the near-in sidelobes, the combined output will drop to zero even though there is noise in each of the receivers! (This is equivalent to any point in the image away from the phase center.)

Another way of stating this is that, if we make an image of an area equal to the primary beam of the individual dishes of such a compact array, then that image is a convolution of the strong point source with the beam of this compact array. Since the beam has little or no sidelobes, the signal and its associated noise appear only at the position of the source.

4.1.2. Snapshot with a dilute array.

If we now separate the dishes to form a dilute array, then the instantaneous "dirty" beam will have many strong and weak sidelobes. A snapshot image of the strong point source made with this array will be a convolution of this dirty beam with the point source, and therefore will look exactly like the dirty beam. At places where no sidelobes appear, the noise will be zero or negligible since $T_A \gg T_R$. The noise in this image is completely correlated, and the correlation function is the dirty beam. There is only one error associated with this image, which is like a scale factor introduced by the source noise. Because the noise is completely correlated between all the antennas, all the measured Fourier coefficients will be scaled in exactly the same way. Another way of stating this is that the source noise is convolved with the instantaneous dirty beam.

If we now 'CLEAN' this image in the usual way, all the responses away from the source will be removed, since they appear exactly at the positions of the sidelobes. The clean image will be noise-free, and the amplitude of the source at the center will be subject to the expected error (Eq. 23–10). This image is therefore equivalent to the one that would be obtained by scanning with a single dish (Section 1.2).

4.1.3. Synthesis with a dilute array.

A synthesis image can be thought of as a combination of a series of snapshots each of which has a slightly different u-v coverage. It is important to note that, in the case of a strong unresolved source that satisfies $T_A \gg T_R$, all the Fourier coefficients measured in a snapshot are scaled in the same way because they are averaged over exactly the same time. If we take another snapshot, displaced in time, then all the Fourier coefficients measured at that time are scaled in the same way, *but the scale factor will be different from the previous snapshot because of source noise.* If we now combine the two snapshots, then we have two sets of measured Fourier coefficients that are slightly shifted in the u-v plane (because of Earth rotation), and are scaled by different factors. The dirty beam of the combined data is a linear addition of the two snapshot beams. Some sidelobes may cancel because of the change in the pattern of the two beams. The dirty image is again a linear addition of the two snapshot images. However, the combined dirty image may no longer look like the combined dirty beams. For example, at positions where the sidelobes cancelled, the contribution from the source may not cancel because of the different scaling factors, caused by the source noise in the two snapshots. We illustrate this below with a specific example.

Consider a one-dimensional array, like the Westerbork Synthesis Radio Telescope. Suppose we make two observations of a strong unresolved source with slightly different arrangements of the telescopes which give us two different beam patterns.

In Observation 1, the telescopes are arranged like a grating as shown in Figure 23–1a. Correlations are made between the movable element M and all the fixed elements, which have a periodic spacing d. The resulting instantaneous beam has a periodic response as shown in Figure 23–1b. Since we are observing a strong unresolved source, the dirty image looks exactly like the dirty beam, as shown in Figure 23–1c. The peak of the dirty image is $S \pm \Delta S_1$, where S is the flux density of the source and ΔS_1 is the error in S at that instant due to the source noise. If we deconvolve Figure 23–1c using the beam in Figure 23–1b, we will get a point source of height $S \pm \Delta S_1$ at the position of the source, and zero everywhere else. In Observation 2, the arrangement of the telescopes is as shown in Figure 23–1d, which gives a different beam, shown in Figure 23–1e. Notice that the alternate positive responses in Figure 23–1b have been replaced by negative responses of the same magnitude in Figure 23–1e, by shifting the movable element by one half-spacing ($d/2$). The dirty image, shown in Figure 23–1f is exactly like the beam in Figure 23–1e, but the peak is now $S \pm \Delta S_2$, where ΔS_2 is the error in S at the time of Observation 2. If we deconvolve the image in Figure 23–1f using the dirty beam in Figure 23–1e, we will again get a point source of height $S \pm \Delta S_2$ at the position of the source, and zero everywhere else. We now combine Observations 1 and 2. The resulting beam, shown in Figure 23–1g, is a linear sum of Figures 23–1b and 23–1e. Alternate responses cancel, since they are equal and of opposite sign in Figures 23–1b and 23–1e. The resulting dirty image is the sum of Figures 23–1c and 23–1f, and is shown in Figure 23–1h. Notice that the alternate responses in the dirty image have not cancelled, since $\Delta S_1 \neq \Delta S_2$. If we now deconvolve the image in Figure 23–1h, using the dirty beam Figure 23–1g, then, in addition to the point source, there will be some residual noise at the positions of the sidelobes.

In a long observation, involving either Earth rotation or movable elements, the instantaneous sidelobes move about in the image and therefore distribute the source noise in the off-source regions as described above. Given that the time

Figure 23–1. **(a)** Arrangement of telescopes for Observation 1. **(b)** Dirty beam for Observation 1. **(c)** Dirty image of a strong unresolved source in Observation 1. **(d)** Arrangement of telescopes for Observation 2. The movable element, M, has been shifted by $d/2$. **(e)** Dirty beam for Observation 2. **(f)** Dirty image of the strong unresolved source in Observation 2. Note that the peak in (f) is different from peak in (c) due to source noise. **(g)** Dirty beam of the combined data, which is a linear combination of (b) and (e). **(h)** The combined dirty image.

taken for a projected separation between two antennas in the array to change by one antenna diameter is small compared to the total observation time, it is reasonable to imagine that all areas of the image would receive more or less equal treatment.

The above experiment illustrates that in synthesis imaging, the instantaneous dirty beams add linearly (addition of voltages), whereas the source noise that appears in the off-source regions due to the sidelobes adds only in quadrature (addition of intensity or power). As a result the distribution of source noise in a synthesized dirty image is not a convolution of the source noise with the synthesized dirty beam. Deconvolution therefore cannot remove this noise.

The above discussion shows that the magnitude of the noise in the off-source region is related to the r.m.s. sidelobe levels of the instantaneous dirty beams. For an array of N elements that are randomly arranged, the r.m.s. sidelobe level is $1/N$ times the peak response. The peak response is proportional to the total collecting area of the array (NA_e) and therefore the r.m.s. sidelobe level is proportional to A_e, the effective collecting area of the individual elements. This follows from Parseval's Theorem. The noise level in the off-source region is therefore given by Equation 23–10. The magnitude of the expected noise can therefore be calculated using the antenna temperature in the individual elements of the array.

4.2. Strong, partially resolved source.

In practice, strong sources (like Cassiopeia A and Cygnus A), for which $T_A \gg T_R$, are not point-like. The characteristics of noise in the image for these cases can be understood if the source is considered to be the superposition of several point sources. The behavior of the noise should be similar to the strong point source case discussed above, i.e., in a snapshot there should be no contribution of the source noise to the off-source region, and in a long-integration synthesis image, the off-source noise is given by the antenna temperature of the individual elements of the array.

4.2.1. Snapshot with a dilute array.

For simplicity, consider a strong double source that dominates the system temperature of the individual elements of the array. As before, a snapshot produces a dirty image that is a convolution of the dirty beam with the double source. The dirty beam appears at the positions of the two strong point sources. The source noise, which is independent for the two sources, will appear as different scale factors for the two beam patterns that appear in the dirty image. If we deconvolve this image, then we should get two point sources with zero noise in the off-source regions. The argument can be continued to an extended source, which can be considered as made up of a number of point sources, each with its own scale factor.

It is important to note that the noise is zero everywhere in the image except at the position of the source, where it is proportional to the flux density of the source. Therefore, if we apply this argument to a fully resolved source that fills the primary beam of the individual antennas, there is noise everywhere in the image, as expected.

4.2.2. Synthesis with a dilute array.

As before, any sequential observation that involves either Earth rotation or movable elements changes the sidelobe pattern as a function of time. The source noise appears everywhere in the image since it is not convolved with the final synthesized beam. It is difficult to derive a general expression for the off-source noise since the noise depends on the structure of the source and the instantaneous u-v coverage of the array. A lower limit to the noise is given by Equation 23–9 by replacing $T_{\rm sys}$ with T_A. An upper limit is given by Equation 23–10. Thus

$$\frac{T_A}{\sqrt{B\tau}} \geq \Delta T \geq \frac{T_A}{\sqrt{N(N-1)B\tau}}, \tag{23–11}$$

where we have assumed $T_A \gg T_R$. The inequality arises because not all the $N(N-1)$ measurements made in any single integration are independent. On the other hand, unlike the strong point-source case, they are not identical. The magnitude of the noise in the off-source region should therefore depend on the structure of the source and on the instantaneous u-v coverage.

4.3. On-source noise.

In all the discussions so far, our main concern has been the off-source noise when observing a strong source that dominates the system temperature. Additional factors that come in to play in determining the on-source noise are (a) the total collecting area of the array, NA_e, and (b) the flux density per synthesized beam area, S_b. The on-source noise differs significantly from the off-source noise when the effective antenna temperature becomes comparable to, or exceeds, the receiver temperature. The effective antenna temperature is calculated using the *total collecting area* of the array, NA_e, and the flux density per synthesized beam area, S_b. For the purpose of determining the on-source noise, a source should therefore be considered as "bright" if it satisfies the criterion

$$T_{A_{\rm tot}} = \frac{S_b N A_e}{2k} \gg T_R\,. \tag{23–12}$$

In a synthesis image of a strong unresolved source, the on-source noise is given by Equation 23–7, where the antenna temperature T_A is calculated using the total collecting area of the array (NA_e). Thus

$$\Delta T \approx \frac{SNA_e}{2k\sqrt{B\tau}}\,. \tag{23–13}$$

When the source is partially resolved, we should use the flux density per synthesized beam, S_b, to calculate the on-source noise. As discussed in Section 4.2.2, we obtain an inequality

$$\frac{S_b N A_e}{2k\sqrt{B\tau}} \geq \Delta T \geq \frac{S_b N A_e}{2k\sqrt{N(N-1)B\tau}}\,. \tag{23–14}$$

The inequality arises because the $N(N-1)$ measurements made in any one integration are not all independent.

4.4. Effect of self-calibration.

Let us consider a strong unresolved source. As noted in Section 4.1.3, successive snapshots made with a synthesis array measure the same values for the Fourier coefficients, apart from a scaling factor that is introduced by the source noise. Because of this changing scale factor, there will be noise in the off-source region in a full-synthesis image. If we determine and remove this scaling factor from each of the snapshots there will be no noise in the off-source region of the synthesized image. This scale factor can be considered as an amplitude gain term for the individual antennas, so if we make an amplitude self-calibration with a point source model these scaling factors will be determined. Thus amplitude self-calibration will remove the off-source noise from the synthesis image of an unresolved source.

For a partially resolved source, there is no single scale factor, since different parts of the source produce independent source noise. Therefore self-calibration will not alter the off-source noise.

4.5. Applications.

One application of the concepts discussed here is in the detection of a weak source within the primary beam in the presence of a strong source that dominates the system temperature. As discussed in Section 4.1.3 the off-source noise will be dominated by the source noise in a full synthesis image. In a snapshot image, however, the off-source noise depends only on the receiver noise (Section 4.2.1). Therefore, if the strong source is removed from every snapshot individually (say by deconvolving and subtracting the 'CLEAN' components from the u-v data), the final synthesis image will have contributions only from the receiver noise. This exercise could be inordinately computer-intensive, but it might be the only method of detecting such a source.

An important result of this analysis is that, when a source is strong (as in Eq. 23–12), the on-source noise can differ significantly from the off-source noise. For a strong source, we cannot use off-source noise estimates to determine the errors in any parameters of the source. By the same argument, in maximum entropy deconvolution algorithms, where an ancillary noise estimate is required in order to make the problem a well-posed one, it may be necessary to specify the noise in each pixel of the image.

In spectral line experiments involving weak absorption or stimulated emission, the sensitivity depends on the strength of the background continuum. If the background source is extended, the sensitivity can therefore be improved by using more compact arrays, since it increases the continuum brightness (flux density per beam area). However, as noted in Section 4.3, the on-source noise also increases with continuum brightness. So, after a point, any apparent increase in sensitivity is exactly offset by a similar increase in the on-source noise. No improvement can, therefore, be expected by using either more compact arrays or a larger collecting area.

References

Christiansen, W. N. and Högbom, J. A. (1985), *Radio Telescopes*, Second Edition, Cambridge University Press (Cambridge, England).

Hanbury Brown, R. A. (1974), *The Intensity Interferometer*, Taylor and Francis, London.

Högbom, J. A. and Brouw, W. N. (1974), "The synthesis radio telescope at Westerbork. Principles of operation, performance and data reduction", *Astron. Astrophys.*, **33**, 289–301.

Rogers, A. E. E. (1976), "Theory of two-element interferometers", in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, pp. 139–157.

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

24. SYNTHESIS OBSERVING STRATEGIES—A "HITCH-HIKER'S GUIDE"

Alan H. Bridle

1. Introduction

Many of the previous lectures emphasize problems that can corrupt or distort the images made by synthesis radio telescopes under real conditions. By now, it may seem to you that so many effects can degrade synthesis images that good results must depend on a mixture of luck and magic. I will try to counteract this by describing an orderly way to choose the parameters of radio synthesis observing to avoid at least *some* of the pathological image defects emphasized earlier. I will suggest how to use the material of Lectures 2, 5, 6, 7, 8, 9, 12 and 13 to choose critical parameters when planning synthesis observations. As there is at least one orderly approach to the problem, my first advice is—"DON'T PANIC!" (Adams 1979).

The general principles of this lecture apply when planning observations with any synthesis radio telescope. Figure 24–1 summarizes the approach that I suggest, as a "decision tree" for continuum observers. Sections 2 to 5 describe details of this tree. Section 6 outlines a similar tree (Figure 24–2) for spectroscopy, and Section 7 discusses calibration strategy. Section 8 describes what (little) can be done while the data are being taken. Many details relevant to the NRAO's VLA are contained in an Appendix. This Appendix also has short sections on preparing VLA observing proposals and on PC software (available from the NRAO) that can assist planning for VLA observing.

2. Resolution, Baseline Range and Observing Frequency

2.1. Resolution θ_{HPBW} – how much is enough?

The first step is to determine the *range* of resolution θ_{HPBW} that is appropriate for the project. You should think about both the minimum and maximum values of θ_{HPBW}.

The lowest resolution (*maximum* value of the synthesized beam width θ_{HPBW}) will be set by your need to separate, or to resolve, astrophysically important features of the region being imaged. Realize however that you can have *too much* resolution if your project depends on imaging *extended* emission. There is no point observing extended emission using such a small beamwidth θ_{HPBW} that crucial features are close to, or below, the r.m.s. noise ΔI_m on the final images. You must therefore estimate the *apparent brightness* (flux density per synthesized beam area) that you expect important extended features to have at the resolution you will use for your final images. You can then determine the highest resolution (*minimum* θ_{HPBW}) that is appropriate for your project by estimating the total integration time t_{int} needed to reach the required brightness

Figure 24–1. Factors Entering Into Continuum Observing Strategy—A Suggested Decision Tree.

sensitivity. As omitting this aspect of the project design can render the final images useless, it is worth restating the principles.

Recall from Lecture 7 that a *point* source with flux density S Jy images with an apparent brightness of S Jy per synthesized beam area regardless of the area Ω_s of the synthesized beam. It follows that, for any synthesis array of identical antennas and receivers, all baselines are equally sensitive to a given point source (apart from the effects of confusion and phase stability). In contrast, as described in Lecture 7, the apparent brightness of an *extended* emission region in a synthesized image depends on the region's detailed structure, on how well the visibility function $V(u, v)$ is sampled by the observations, and on the weighting and tapering functions D_k and T_k applied to the data when imaging (Lecture 6). When deciding on an observing strategy, it usually suffices to assume that:

(a) an extended region with uniform true brightness I Jy per arcsec^2 will be imaged with an apparent brightness $\approx I\Omega_s$ Jy per synthesized beam area, and

(b) the *final* synthesized beam will be a Gaussian 'CLEAN' beam, so that its area Ω_s will be $\approx 1.13\theta_1\theta_2$ arcsec^2, where θ_1 and θ_2 are the major and minor half-power widths of the Gaussian in arcsec.

If the r.m.s. noise on the image is ΔI_m Jy per synthesized beam, the signal-to-noise ratio of such extended emission on the image will be $\sim I\Omega_s/\Delta I_m$, which increases as the synthesized beam area Ω_s. You must ensure that you do not observe at such small values of Ω_s that interesting extended structure is undetectable, given the total integration time $t_{\rm int}$ available and your choice of the IF bandwidth $\Delta\nu$ (see Sections 3 and 4 below).

There are circumstances however when enhanced resolution helps you to detect interesting features—for example, when searching for pointlike "hot spots" or linear "jets" in more diffuse emission such as large scale "lobes". While the flux density per synthesized beam of two-dimensional emission is roughly proportional to the beam *area* Ω_s, that of linear emission is proportional to the beam *width* $\theta_{\rm HPBW}$, and that of a point source is *independent* of beam size. These dependencies allow compact structure that is embedded in, or confused with, more extended emission to be recognized most easily on high-resolution images.

Note that it is also important to avoid unnecessarily high resolution (long baselines) in detection experiments. Although the *theoretical* sensitivity to a point source is independent of the size of the array (apart from the effects of confusion), the phase fluctuations produced by atmospheric irregularities will be greater on longer baselines (see Lecture 5, Section 4.3). It is therefore more difficult to reach the theoretical sensitivity by integrating coherently between calibrations when using long-baseline arrays. This is especially true at high frequencies, where the phase stability depends critically on conditions in the troposphere over the array. The severity of ionospheric or tropospheric phase fluctuations varies from site to site, and at any one site with the "weather" from hour to hour, from day to day, and from season to season. It is generally true however that observations longer than about an hour with $\theta_{\rm HPBW} < 1''$

are often corrupted by atmospheric phase fluctuations and that observations with $\theta_{\rm HPBW} < 0\farcs1$ *always* are. The most powerful tool for dealing with these corruptions is self-calibration (Lecture 9). This technique is rarely applicable to detection experiments, however.[1]

2.2. Largest angular scale $\theta_{\rm LAS}$.

The next step in planning your observations involves thinking about $\theta_{\rm LAS}$, the largest angular scale of structure that you must sample well to produce a useful final image. $\theta_{\rm LAS}$ will be the angular diameter of the most extended structure that you must image accurately—usually the diameter of the most extended component of astrophysical interest in your target. (Do not confuse it with $\theta_{\rm max}$, the required field of view, which is discussed below—when observing a target $10''$ in extent in the presence of a point confusing source $1'$ away, you would set $\theta_{\rm LAS} = 10''$, not $\theta_{\rm LAS} = 1'$.) $\theta_{\rm LAS}$ will set the *shortest* baseline length that must be present in your data for the resulting images to be useful.

The visibility of a circular Gaussian source with FWHM $\theta_{\rm LAS}$ falls to half of its peak amplitude at a normalized baseline length of $91{,}000/\theta_{\rm LAS}$ wavelengths, if $\theta_{\rm LAS}$ is in arc seconds. You will not be able to image a two-dimensional structure of size $\theta_{\rm LAS}$ well unless *projected* baselines at least this short are present in the data. Depending on the density of tracks in the inner u-v plane, and on the sensitivity per baseline, more conservative criteria may be needed—halving the above estimate gives a baseline on which the visibility falls to about 85% of its peak. Note also that the detailed array geometry, target declination and hour-angle range determine how baselines are foreshortened by projection. The most reliable approach is therefore to look at u-v plots for your expected antenna layout, declination, and hour-angle range,[2] to ensure that your planned short-baseline coverage meets the requirements imposed by $\theta_{\rm LAS}$.

The factors affecting the choice of resolution cannot be estimated reliably in advance if the source structure is poorly known. If you are not sure what your source will look like, the safest strategy is to guess on the side of low resolution when designing an initial observation. A "diagnostic" low resolution image may tell you the source's total angular extent, and warn you about surrounding emission. This information will let you optimize the parameters for a more time-consuming high resolution study. It is also easier to justify reobserving a detected emission region at high resolution than to justify reobserving at low resolution what previously appeared to be empty sky!

2.3. Choice of frequency ν_0 at given resolution $\theta_{\rm HPBW}$.

Continuum observers will choose the observing frequency *at a given resolution* by balancing astronomical criteria against instrumental performance. A high frequency may be necessary for polarimetry because Faraday effects decrease with increasing frequency (degrees of linear polarization are generally higher at higher frequencies and electric vectors lie closer to their intrinsic position angles). Polarimetric data are also less corrupted by the Earth's ionosphere at

[1] Unless they are unusually successful, so that all detected sources are strong!
[2] This can readily be done for any antenna array using the program 'UVSIM' in the NRAO's Astronomical Image Processing System.

high frequencies. Studies of Faraday effects (e.g., depolarization) in continuum sources may demand low observing frequencies, however. The spectral index of the continuum also influences the choice of frequency—optically thick thermal emission may be easier to detect at high frequencies whereas transparent synchrotron sources will be easier to detect at low frequencies.

If a continuum project has frequency agility, the choice of ν_0 also involves the receiver sensitivity and antenna performance as a function of frequency. High frequency observations are often limited by worsening receiver noise performance, antenna efficiency and pointing accuracy (as a fraction of the primary beam size) with increasing frequency. Low frequency observations are limited by interference at many radio observatories.

a) Interference.

The detailed choice of observing frequency, especially below 2 GHz, is often constrained by known sources of radio frequency interference (RFI). For example, U.S. frequency allocations in the L Band around 20 cm include aeronautical radio navigation (including global satellite networks), meteorological aids, and fixed and mobile use. Many of the possible external interfering signals are time variable, so freedom from external interference can never be guaranteed at L Band outside the protected radio astronomy bands. The protected band from 1400 to 1427 MHz is undesirable for some continuum observations, however, because galactic neutral hydrogen emission can increase the system temperature significantly in this band.

External interfering signals are partially rejected by interferometers (e.g., Thompson 1982a) because only the component of the signals that (a) varies at the sidereal fringe rate, and (b) correlates with the correct delay, will affect the output. In practice, this combination makes low frequencies and short-baseline synthesis arrays more sensitive to RFI than high frequencies (where radio frequency band allocations are less congested) and long-baseline arrays.

Data averaging reduces the response to interfering signals by the factor $\mathrm{sinc}(\nu_f \tau_c)$ where ν_f is the natural fringe frequency and τ_c is the time over which the data are averaged when computing the visibility in each u-v cell before Fourier transforming. As $\nu_f = \omega_e u \cos\delta$, where ω_e is the angular velocity of the Earth's rotation and δ is the declination, fringe rate rejection is least for observations near the celestial poles and for baselines near the v-axis, where fringe frequencies are near zero. In a given RFI environment, long-baseline arrays collect more data at large values of u than do short-baseline arrays, so long-baseline arrays are less affected by RFI (unless the RFI is strong enough to degrade the noise performance of the individual receivers). Short-baseline arrays may also suffer more from locally-generated RFI or from "crosstalk" (correlations between signals radiated by the electronics of one antenna directly into the feed or the electronics of another).

Delay rejection is a more complex function of the azimuth of the baseline and the position of the interfering source, as described in detail by Thompson (1982a). Its effects cannot be stated as concisely as those of fringe rate rejection and are not a simple function of location in the u-v plane. Delay rejection is not usually significant for narrow-band interfering signals.

RFI may always be edited from the data after they are taken (with particular attention to data falling near the v-axis), but careful choice of observing frequency can minimize both the effort necessary for editing and the consequent loss of data.

b) Imaging at more than one frequency.

When studying spectral index, Faraday rotation or depolarization properties of extended sources, you should attempt to minimize uncertainties arising from differing resolutions at the different frequencies. It is important to try to "scale" the distributions of baseline lengths in *wavelengths* that are used for the different observations. Ideally, you would scale the detailed distribution of antenna spacings but it is often difficult to scale more than the inner and outer baseline limits and the tapering and weighting functions (especially when observing with fixed arrays, as in VLBI). You should also make the hour-angle ranges of the observations at different frequencies as similar as possible.

Note that although using such "scaled arrays" *optimizes your chances* of measuring frequency-dependent properties of a source accurately, it cannot *guarantee* success. Even scaled arrays may have significantly different relative sensitivities to different scales of structure if the structure changes radically with frequency, e.g., if there are large spectral index gradients in either the total or polarized emission. You must also be careful when interpreting the final images if the databases at different frequencies are differently affected by missing antennas or by bad data. In many cases, the reliability of inter-frequency comparisons depends on how well the deconvolution algorithm (Lecture 8) interpolates unsampled parts of the u-v plane, even when scaled arrays have been used.

Observing at several adjacent, but not necessarily contiguous, frequency bands can improve the u-v coverage, and so help you to reconstruct complex structures with high fidelity, by "bandwidth synthesis" (see Lecture 12). For this application, observing frequencies should be spaced to minimize the sizes of residual "holes" in the u-v coverage. For example, if the largest "holes" $\Delta u, \Delta v$ in the coverage at one frequency are about 10% of the corresponding radii $\sqrt{u^2 + v^2}$, data from two observing frequencies separated by 5% might be combined.

3. Field of View Restrictions

Once you have settled on the resolution θ_{HPBW}, baseline range and observing frequency ν_0 for your program, the next decision (see Fig. 24–1) involves IF bandwidth $\Delta\nu$ and averaging time τ_{a}. These must be consistent with the field of view requirements of your program. You must therefore consider the radius θ_{max} (from the center of the field of view) over which you require the data to be minimally distorted by chromatic aberration and time-average smearing as discussed in Lectures 2, 12 and 13.

3.1. IF bandwidth $\Delta\nu$.

An unsuitable choice of the IF bandwidth may lead (a) to irrecoverable distortion of the image by chromatic aberration if the bandwidth is too great (see Lectures 2, 12 and 13) or (b) to unacceptably low sensitivity if it is too small.

As discussed in detail in Lecture 13, images made with finite bandwidth are degraded away from the delay tracking center by radial smearing and reduction in amplitude of the point source response. The first step in choosing the IF bandwidth for continuum work is therefore to ask over what field radius $\theta_{\rm max}$ (arcsec) you require that these chromatic aberration effects be limited to broadening the beam by $< n\%$ or to reducing the amplitude of a point source by $< m\%$. Then obtain the corresponding value $\beta_{\rm max}$ of the normalized parameter β, either from the graphs shown in Figures 13–1 and 13–2, or from Equations 13–19, 13–24 and 13–29. The maximum allowable IF bandwidth $\Delta\nu_{\rm max}$ (MHz) consistent with these constraints is then given by

$$\Delta\nu_{\rm max} = \frac{\beta_{\rm max}\nu_0\theta_{\rm HPBW}}{\theta_{\rm max}}\,, \tag{24–1}$$

where ν_0 is your observing frequency in MHz and $\theta_{\rm HPBW}$ is the half-power beamwidth in arcsec at which you expect to make your images. Unless you are prepared to relax your smearing/attenuation criterion slightly, select the closest allowed bandwidth that is *narrower* than $\Delta\nu_{\rm max}$. If you are prepared to relax it, choose the closest *wider* bandwidth.

Your choice of $\theta_{\rm max}$ may be determined by the need to image an extended structure with minimal distortion, or by the need to include a strong confusing source in the minimally-distorted field of view. The latter need arises because you may wish to subtract or deconvolve a confusing source's sidelobes from the interesting region. The value of $\theta_{\rm max}$ will always be greater than, or about equal to, the value of $\theta_{\rm LAS}$ used earlier when selecting the range of baselines. In general, choose the delay and pointing center to minimize the required $\theta_{\rm max}$ for your observations. When using a wide field to include a confusing source, consider displacing the delay center away from the target towards the confusing source. This will avoid the use of unnecessarily narrow bandwidths (and thus of unnecessarily low sensitivity). If the field is *dominated* by a strong point source (more than ten times brighter than other structure), *this* source should be placed near the delay center and pointing center when high dynamic range is required. This strategy will minimize the total distortion of the image resulting from bandwidth, pointing, averaging time and computational effects (such as u-v truncation) involving the strong source (for more details, see Clark 1981).

For point source detection experiments the above criteria will normally select the widest available bandwidth, unless the search position is exceptionally inaccurate or the field is known to be highly confused. Chromatic aberration affects observing strategy less critically at higher frequencies for several practical reasons—fractional bandwidths ($\Delta\nu/\nu_0$) are often smaller; the usable field of view is limited by the primary HPBW of the antennas for the narrower fractional bandwidths; receiving systems are noisier and antennas are less efficient (so sensitivity, not distortion, limits image quality).

When deciding on the value of $\theta_{\rm max}$ that is appropriate for an image of an extended source, consider the detectability of the extended emission *at the* *resolution you will be using for your images* (see Sections 2 and 4). There is no point ensuring that extended structure is not smeared radially by the bandwidth effect if low signal-to-noise on the same structure introduces uncertainties larger than the bandwidth distortions. As the signal-to-noise on extended emission itself depends on the choice of IF bandwidth, this calculation may need to be iterated until a suitable compromise is reached.

3.2. Visibility averaging time $\tau_{\rm a}$.

It is important to choose the "on-line" averaging, i.e., averaging applied to the data before they are written to an initial storage medium, carefully. The data will be averaged relative to a particular phase tracking center and this on-line averaging cannot be undone. Any excess averaging done off-line, after the data have been archived, can be undone later, though it is obviously inefficient to do this.

The effects of finite averaging time $\tau_{\rm a}$ were outlined in Lectures 2 and 12 and discussed in detail in Lecture 13. As $\tau_{\rm a}$ is increased, phase winding of a point source at radius θ from the phase center both smears and attenuates the synthesized response to such a source. The effect is worst on a given baseline when Earth rotation moves the source perpendicular to the fringes associated with that baseline, but is zero when the feature moves parallel to the fringes. The magnitude of the effect therefore depends on baseline orientation, hour-angle and declination, as described in Lecture 13 by Equations 13–31, 13–35, 13–36 and 13–37. For an array observing a point source near the north celestial pole, the *average* reduction in amplitude can be estimated from Equations 13–41 to 13–43, provided the reduction is small. Restating those results, the average reduction in amplitude of a polar point source θ from the phase tracking center is given by

$$\overline{\langle R_\tau \rangle} = \frac{I}{I_0} = 1 - \eta \left(\frac{\theta}{\theta_{\rm HPBW}} \right)^2 \tau_{\rm a}^2 \,, \tag{24–2}$$

where $\eta = 1.05 \times 10^{-9}$ for uniform square u-v coverage, $= 1.08 \times 10^{-9}$ for uniform circular coverage and $= 1.22 \times 10^{-9}$ for circular, Gaussian tapered, coverage. $\tau_{\rm a}$ is in seconds.

The choice of $\tau_{\rm a}$ balances several issues. First, you must decide what amplitude reduction R_τ due to time-average smearing will be acceptable at $\theta = \theta_{\rm max}$. Then estimate the corresponding $\tau_{\rm a}$ from Equation 24–2 using appropriate values of η and $\theta_{\rm HPBW}$. If sensitivity or hardware considerations force you to choose an IF bandwidth that is wider than the $\Delta\nu_{\rm max}$ given by Equation 24–1, estimate $\tau_{\rm a}$ so that the effects of time-average smearing at $\theta_{\rm max}$ are slightly less than those of chromatic aberration. (See the Appendix, Section A4, for an example.)

Ideally, you would observe with $\tau_{\rm a}$ set to the value estimated from Equation 24–2 unless this exceeds the expected coherence time $\tau_{\rm atm}$ for the atmospheric phase fluctuations on the longest baselines, or unless it is too long to let you edit the data satisfactorily. In either of those cases, you would observe with a shorter $\tau_{\rm a}$. You would later average the edited, calibrated data to the value estimated from Equation 24–2, to minimize the data volume and the image processing time. In the real world, constraints on computing resources (storage space and CPU time) may force you to use longer averaging times than would be desired purely for image quality.

4. Total Integration Time $t_{\rm int}$

Once you have determined the IF bandwidth $\Delta\nu$ from the field of view criteria, the next step in the decision tree (Fig. 24–1) is to estimate the total on-source integration time $t_{\rm int}$ required for given sensitivity on your final image. Here you will use the expression for the r.m.s. noise ΔI_m on an image made with an N-antenna array:

$$\Delta I_m = F_{\rm w} \Delta S \Bigg/ \sqrt{\frac{nN(N-1)}{2} t_{\rm int} \Delta\nu} \ , \qquad (24\text{–}3)$$

where n is the number of independent IFs contributed to the image per antenna ($n = 2$ for images of Stokes I from two orthogonal polarization states at one sky frequency, or for images of $P = \sqrt{Q^2 + U^2}$ at one sky frequency), $t_{\rm int}$ is in seconds, and $\Delta\nu$ is in MHz. In the numerator, $F_{\rm w} = 1.0$ for natural weighting and is > 1 for other weightings (the value depending on the gridding and on the baseline distribution—see Lecture 7 for details). ΔS is the single-interferometer sensitivity per second per MHz of IF bandwidth.

The sensitivity goal will be determined by (a) the significance level you need for a detection to achieve your astronomical goals, and (b) whether the interesting emission is extended (see Section 2.1 above). If you are doing polarimetry, calculate the sensitivity from the expected *polarized* intensity, not total intensity.

If the first estimate of $t_{\rm int}$ is significantly greater than 12 hours, consider carefully whether your choices of frequency and baseline range are optimal. You may wish to re-enter the decision tree with different starting parameters before considering the proposal planning further.

If you estimate $t_{\rm int} < 4$ hours, your observing strategy should be determined by the need for dynamic range and by whether you can merge observations of several sources into one program. If you require high dynamic range, or wish to image an extended structure, with less than a full synthesis you should sample the u-v plane as uniformly as possible. This can usually be done satisfactorily by distributing the observing over several short (e.g., $\sim$ 10-minute) scans spaced equally through the available hour-angle range. Note however that the dynamic range achieved in a given observation also depends on atmospheric and ionospheric conditions, on the elevation angle range, and on your calibration strategy (Section 7 below), as well as on the u-v coverage.

Two-dimensional arrays with many antennas may have good enough instantaneous u-v coverage to make a "snapshot" mode attractive for observations of strong, compact sources when the total integration time required is small. The Appendix (Section A7) discusses the advantages and disadvantages of snapshot mode.

5. Confusing Sources

Confusion may have two effects on the interpretation of a synthesis image:

(1) the r.m.s. fluctuation level on the image may be degraded by sidelobes or by aliasing of confusing sources, and

(2) an extraneous radio source may be mistaken for the target in a detection experiment, or considered part of the structure of an extended feature.

The best way to handle confusion is to *learn about it* in the context of your particular target field. Before you observe, consult published radio surveys, or low resolution images of the field around your target, for possible confusion. If you know you will be observing near a bright confusing source, you may consider either of two strategies for reducing its effects on your final images.

If the confusion is relatively nearby, plan to look at it—to make wide-field images containing both the target and the confusing source. Then subtract or 'CLEAN' the confusing source and its sidelobes from the interesting region. The ungridded subtraction technique[3] (Lecture 8, Section 1.3) helps this strategy, as only the parts of the wide field that contain significant emission need to be computed and 'CLEAN'ed. This approach can work well if the confusing source is only one or two times as far from the target as the radius of the field of view that you would otherwise have imaged. If the confusing source is strong, consider displacing the delay tracking center away from the target toward the confusion, to minimize distortions of the response to the confusing source by chromatic aberration and other effects.

If the confusion is relatively distant, try *not* to look at it—choose your IF bandwidth and the delay, phase and pointing centers to minimize the response to the confusing source. Note that because the point source attenuation produced by chromatic aberration and time averaging increases with baseline length $\sqrt{u^2 + v^2}$ (see Lectures 12 and 13), these effects do not filter confusing sources from the short-baseline data as effectively as they do from long-baseline data. If a distant confusing source still dominates the data after attenuation by the primary beam, it may therefore produce wide-angle "ripple" in the final image. In such cases, the pointing center should be chosen to minimize the response to the confusion rather than to maximize the response to the target.

A difficult case arises when the response to the confusing source is strong even after adopting this stratagem. Variable pointing errors and (for altitude–azimuth antenna mounts) the rotation of the primary sidelobe pattern of the antennas on the sky may then make the confusing source appear to vary throughout the observation. It is hard to make images with high dynamic range in these circumstances (see also Lecture 12, Section 2.1). It is worth recognizing this difficulty before observing, if only because you might be able to select an alternative target that is equally interesting.

If a confusing source lies in the target field itself, no advance precautions are necessary, as the confusing source and its sidelobe pattern can be deconvolved

[3] coded in the NRAO Astronomical Image Processing System as the program 'MX'.

from the image as part of normal data reduction. Note however that it is important to *look* for confusion *after* taking the data, e.g., by making a heavily-tapered low-resolution image covering as much sky as possible around your target. It is particularly important to do this before proceeding with computer-intensive processing such as deconvolution, or self-calibration at full resolution. Making low-resolution images to search for confusing sources early in your image processing may (a) improve the models used for self-calibration, (b) reveal unexpected large-scale structure around your target (and so help to define 'CLEAN' deconvolution windows) and (c) lead to unexpected discoveries.

In detection experiments, confusion may make the interpretation of a positive detection questionable if a source is detected near, but not at, the target position. Source counts (e.g., Section A9 of the Appendix) should be used to estimate the probability that the detected source lies in the image by chance.

6. Spectral Line Observing

Spectroscopists have to balance slightly different constraints than those confronting continuum observers. Figure 24–2 summarizes an alternate decision tree appropriate for spectral line observing.

6.1. Observing frequency.

Many spectroscopic experiments have little choice of observing frequency. The line frequency is set by the chemical species and transition relevant to the astronomical problem. The observing frequency is then determined by correcting the line frequency for the systemic radial velocity and for the various motions of the Earth (see Lecture 17, Table 17–1). For the experiment to be viable, the corrected frequency must, of course, lie within the tuning range of the receivers and feeds.

If the chemical species has many radio-frequency transitions (e.g., recombination lines), instrumental performance and astronomical criteria may again be combined. For such projects, the instrumental criteria are similar to those for the continuum: receiver sensitivity, antenna efficiency, pointing accuracy. The astronomical criteria may include whether different transitions will be in emission, or absorption, or both.

The last factor to be considered is radio frequency interference (RFI), as described in Section 2.3*a*. Some spectroscopic observations, especially those at frequencies below 2 GHz and those of objects whose red shifts move the observing frequency out of a protected band, are simply not possible because of RFI.

6.2. IF bandwidth.

Two IF bandwidths enter into the observing strategy for spectral line work. The first is the *total* IF bandwidth $\Delta\nu_{\rm tot}$ that is spanned by all the line channels. $\Delta\nu_{\rm tot}$ determines the *range* of velocities included by the experiment, which should span not only the interesting parts of the line profile, but also an adequate number of line-free channels (allowing for any band-edge effects in autocorrelation spectrometers). The second is the individual *channel* bandwidth $\Delta\nu_{\rm ch}$. This sets the velocity *resolution* of the experiment. It also determines the chromatic aberration effects for the individual channel images. As indicated in Figure 24–2,

Figure 24–2. Factors Entering Into Spectral Line Observing Strategy—A Suggested Decision Tree.

the choices of $\Delta\nu_{\rm tot}$ and $\Delta\nu_{\rm ch}$ are usually tightly coupled by the design of the spectral line correlator, and velocity resolution and velocity span must often be traded against one another.

The need for velocity resolution usually drives the *channel* bandwidth to a small value so that individual channel images are negligibly smeared by chromatic aberration. This does not mean, however, that spectroscopists can ignore chromatic aberration in synthesis imaging. The *synthesized beam* pattern is different for every channel image when the channels are gridded separately with their correct frequencies. Any data reduction schemes that do not separately deconvolve the synthesized beam from each of the individual channel images may therefore "rediscover" effects related to the frequency dependence of the beam. As Lecture 18 describes, such separate deconvolution of all the channels is not always a good strategy. If you estimate the continuum distribution by summing the "dirty" images over a range of line-free channels, or if you difference two "dirty" line channel images, the variation of the synthesized beam with frequency may become apparent. If you try to avoid this effect for some channels by artificially assigning them the same frequency (to make the "dirty beam" the same for all images), you will restore the chromatic aberration corresponding to the widest channel separation $\Delta\nu_{tot}$ used in such processing.

6.3. Averaging time $\tau_{\rm a}$.

Ideally, the averaging time $\tau_{\rm a}$ for spectroscopy will be determined by considering (a) the distortion produced by time-average smearing (Equation 24–2), and (b) the atmospheric coherence time $\tau_{\rm atm}$. Because narrow-band channel images are relatively free from chromatic aberration, it rarely makes sense to choose $\tau_{\rm a}$ for high-resolution spectroscopy by matching time-average smearing to chromatic aberration, as in some continuum experiments.

In practice, however, observations with narrow channel bandwidths may also require long *total* integration times $t_{\rm int}$ to reach sufficient signal-to-noise in the individual channel images. This interacts with the choice of $\tau_{\rm a}$, because a long averaging time may be needed simply to restrain the number of independently recorded visibilities. If $\tau_{\rm a}$ is limited by the data volume, Equation 24–2 should be used to check the distortion at the edge of the interesting field. If a $\tau_{\rm a}$ that keeps the data volume manageable also produces significant distortion, you may have to balance using (a) long averaging times $\tau_{\rm a}$ (image distortion), (b) wider channel bandwidths $\Delta\nu_{\rm ch}$ (poorer velocity resolution), (c) fewer channels $\Delta\nu_{\rm tot}/\Delta\nu_{\rm ch}$ (poorer velocity span), or (d) fewer antennas N (poorer image quality) to restrain the data volume ($\propto N(N-1) \times (t_{\rm int}/\tau_{\rm a}) \times (\Delta\nu_{\rm tot}/\Delta\nu_{\rm ch})$).

6.4. Sensitivity and total integration time $t_{\rm int}$.

Spectroscopists may need to consider separately two different sensitivities—the sensitivity to spectral features in the individual line channels and the accuracy with which the continuum can be estimated from the line-free channels. The need for individual channel sensitivity normally dominates, however. The calculation uses Equation 24–3, with $\Delta\nu_{\rm ch}$ replacing $\Delta\nu$. It may also be appropriate to express the sensitivity in terms of brightness temperature by converting from mJy per beam area to temperature units assuming a Gaussian beam of area

$\Omega_s = 1.13\theta^2_{\mathrm{HPBW}}$.

Use of narrow channel bandwidths $\Delta\nu_{\mathrm{ch}}$ often implies that sensitivity, rather than u-v coverage, considerations determine t_{int} for high-resolution synthesis spectroscopy, so that observations > 12 hours long may be needed. Spectroscopic observations may also need long integrations on the calibration sources to achieve enough sensitivity.

6.5. Time of day.

Because spectroscopic experiments commonly use compact arrays and narrow bandwidths, they are more prone than continuum observations to disruption by sidelobe responses to solar radio emission. (Both short baseline lengths and narrow channel bandwidths reduce the immunity from far-out sidelobe responses that stems from bandwidth smearing.) This problem is especially severe at low frequencies and at times of enhanced solar activity. Below 2 GHz, night-time observations may be obligatory for many spectroscopic experiments when the Sun is active.

7. Calibration Strategy

The basic issues to be decided by the observer are: (a) what system-wide parameters will be determined satisfactorily by special purpose calibrations before the observing and (b) what time-dependent calibration parameters will be monitored, and how often, during the observing itself. The first decision will vary from observatory to observatory and from project to project. It must be approached by reviewing all the calibration steps described in Lecture 5, and asking how well each calibration parameter will be known in advance relative to the needs of the project. I concentrate here on the second issue, which itself has two main parts—how often to monitor the instrument and the atmosphere by observing reference sources, or "calibrators", and how close the calibrators should be to the target(s). Your strategy will depend on whether you attempt to calibrate only the instrumental fluctuations, or these fluctuations plus the apparent amplitude and phase variations introduced by the ionosphere and troposphere (again, see Lecture 5 for details). Any calibration that monitors atmospheric fluctuations will, of course, calibrate instrumental fluctuations also.

7.1. Instrumental calibration.

The instrumental calibration should (a) detect grossly malfunctioning antennas so that faults can be corrected while the observations are in progress, and (b) monitor the amplitude and phase stability of the instrument rapidly enough that changes can be corrected by interpolation throughout the run. In a well-designed array, most instrumental fluctuations (apart from phase jumps) are slow, so that observations of an unresolved strong calibrator every few hours or more will monitor the instrument adequately. Bear in mind however that if the instrumental calibration detects a discontinuity such as a phase jump, you may have to discard all the data between consecutive calibration observations for the antenna-IF in which the jump occurred (unless the source being imaged is strong enough that the precise time of the jump can be located in the source data—in which case you can probably self-calibrate it anyway).

Calibrators for purely instrumental monitoring should be chosen primarily for their strength rather than for closeness to the target(s).

7.2. Atmospheric calibration.

It is more important, and much more difficult, to calibrate amplitude and phase fluctuations that result from changes in the propagation properties along the atmospheric path to the target. Unfortunately, *no calibration based on observations of a phase reference source that is not in the same isoplanatic patch as the target can be guaranteed to improve the data quality.* This does not mean that attempts to calibrate atmospheric fluctuations using distant reference sources are a waste of time, but you must recognize that such "external" calibration may or may not be successful.

a) External calibration.

If the target and calibrator are separated by an angle greater than that subtended by the atmospheric cells responsible for the amplitude and phase variations, the fluctuations seen in the calibrator data may not correlate with those in the target data. Corrections interpolated from the calibrator observations into the rest of the data under these circumstances may *worsen* the atmospheric amplitude and phase noise on the target by a factor of $\approx \sqrt{2}$. At the other extreme, if the target and calibrator are typically within the *same* isoplanatic patch, the fluctuations observed in the calibrator will faithfully track those occurring in the target. Amplitude and phase corrections interpolated into the target data from the calibrator data in time series may then greatly improve the quality of the final image. The basic problem is that the scale size of the isoplanatic patch will vary from day to day and even from hour to hour (as a function of the "weather" and of the elevation of your target above the horizon). It is therefore difficult to judge how reliable amplitude and phase referencing from a distant calibrator may be before the observations begin.

If you cannot, or do not wish to, rely on self-calibration to remove atmospheric effects from your data, you should choose external calibrator(s) as close as possible to the target(s), ideally within a degree or so of them. You must then hope that the amplitude and phase stability found in the calibrator data meet the needs of your experiment, and that the scale of the isoplanatic patch is typically greater than the calibrator–target separation.

If the within-scan *and scan-to-scan* amplitude or phase fluctuations on a calibrator a degree or two from your target are small (less than 10% or 20°), it is unlikely that much larger fluctuations are occurring on your target and correcting the rest of the data using long-term averages of the calibrator phases is quite likely to improve matters.

If you see large, rapid fluctuations on the calibrator, you are in trouble, which may or may not be mitigated by correcting the rest of the data by local (point-to-point) interpolation in the fluctuations. Because such interpolation may indeed make matters worse, you might then try making images (a) after long-term averaging, and (b) after local interpolation, of phase corrections from the calibrators. You can then judge empirically which gives better final images—using the final dynamic range, r.m.s. noise level, and/or any prior knowledge about the source to make your judgment.

Deleting data from some or all baselines during periods of unusually bad phase stability will usually improve the quality of images made by external calibration. Imaging with less data but with good amplitude and phase stability can give better results than imaging with much poorly-calibrated data—if the synthesized beam is closer to the theoretical "dirty" beam, deconvolution algorithms work better, increasing the dynamic range of the images. Note that tapering the u-v data down-weights the visibilities from the longer baselines where phases are less stable. Be ready to sacrifice resolution in favor of forming the theoretical beam more closely when the phases are unstable, if your astronomical goals can be met at lower resolution.

Observers doing detection experiments are usually forced to use external calibration. Fortunately, they do not usually require such high dynamic range (and such good phase stability) as observers imaging bright emission complexes. The loss of point source response produced by poor phase stability in a detection experiment can be estimated by calibrating with a long (e.g., $>$ 2 hour) interpolation of the gains, then imaging the calibrators and comparing their apparent flux densities in these images with their assumed values.

b) Self-calibration.

The most reliable method for removing atmospheric fluctuations from the data is self-calibration, *if the target is strong enough.* Self-calibration should be used if the signal-to-noise on the target is high enough to determine the gains in the typical coherence time $\tau_{\rm atm}$ of the atmospheric phase screen, using baselines involving all antennas. For an array of N elements, the r.m.s. noise in the determination of an individual antenna phase gain by self-calibration on a point source of flux density $S_{\rm sc}$ Jy will be

$$\phi_{\rm rms} \approx \frac{65\Delta S}{S_{\rm sc}\sqrt{(N-2)\Delta\nu\tau_{\rm sc}}}\ \text{degrees}\,, \tag{24–4}$$

where ΔS is the sensitivity in Jy per second per MHz of bandwidth, and $\tau_{\rm sc}$ is the averaging time used for the gain solutions. Unless $\phi_{\rm rms} \ll 60°$ when $\tau_{\rm sc} = \tau_{\rm atm}$, self-calibration will not improve the data.

There are several reasons why external calibration is useful even if you know you will be able to self-calibrate your final images. External calibrators will provide flux-density and position scales for self-calibrated images (on which this information will otherwise generally be lost). Observations of the time scale of the phase fluctuations on a phase calibrator near your target can also be used to estimate the coherence time $\tau_{\rm atm}$ of the atmosphere during your observations. This will enable you to select a suitable averaging time $\tau_{\rm sc}$ for the self-calibration: it should be $\leq \tau_{\rm atm}$. Such observations may also tell you that some parts of your data were obtained under more stable atmospheric conditions than others; the "good" parts may yield a better initial model of your target that helps self-calibration of the whole data set to converge quickly.

Fortunately, the class of source for which images of high dynamic range are most important is also the class for which self-calibration is most likely to work well—namely, those with weak extended structures around bright small-diameter components, as discussed in Lecture 16. There is however a range of flux densities and structural complexities over which self-calibration cannot be guaranteed to work in typical atmospheric coherence times $\tau_{\rm atm}$, and for which external calibration therefore remains essential.

7.3. Position calibration.

The phase referencing that is done during external calibration refers the absolute position scale of the images to the mean position scale defined by the adopted positions of the external phase calibrators, if the antenna co-ordinates are well determined. No further positional calibration is required for routine (non-astrometric) synthesis imaging. Note however that *self-calibration* fundamentally has no absolute phase reference and therefore produces images whose position scale is whatever was specified in the input model.

7.4. Flux density calibration.

The flux density scale is usually set by observing an unresolved reference source whose flux density is believed to be stable and whose intensity has previously been measured on an absolute scale. As there are few such sources that remain unresolved to long-baseline arrays, it is rarely possible to use a primary flux density calibrator also as the primary phase calibrator. Good phase calibrators are strong and unresolved, so are likely to vary with time. The flux densities of phase calibrators must usually be determined at the time of the observations by measuring their amplitude ratios to standard flux density calibrators, often using only a short-baseline subset of all the antennas in the array. Note that self-calibration has no absolute amplitude scale and therefore produces images whose flux density scale is whatever was specified in the input model.

7.5. Polarization calibration.

The quantities to be calibrated are the individual antenna cross-polarization terms and the position angle scale (including any time-variable ionospheric Faraday rotation). The design of this calibration depends on whether the antennas are on equatorial or altitude–azimuth mounts, and on the feed polarizations. The VLA case is described in the Appendix.

7.6. Bandpass calibration.

Spectral line observers must also calibrate the instrumental bandpass, often for both amplitude and phase. The amplitude calibration may sometimes be estimated by dividing the cross-correlation spectra by the geometric mean of the autocorrelation spectra of the relevant antennas. This approach is unsuitable however if there are lines (or interfering signals) in the passband that are strong enough to show up in the single-antenna spectra. It is inadequate if the phase variation across the band is significant relative to other uncertainties, for example those imposed by the signal-to-noise ratio.

The most complete approach to bandpass calibration is to observe a strong line-free continuum source to find the bandpass characteristics (amplitude and phase) of each antenna-receiver combination. In a well-engineered system, these characteristics will be only slow functions of time, so the calibration need not be repeated often. Notice however that the total time needed for the bandpass calibration is set by the sensitivity required to measure the gain *in each channel* as accurately as needed for the required *spectral* dynamic range. Projects that require high spectral dynamic range may therefore have to invest a large fraction of the total observing time in bandpass calibration. See Lecture 18 for details of bandpass calibration strategy.

8. Stormy Weather and What to Do About It

Some observing programs have frequency agility. For these, it may be practical to adjust the strategy to take account of the weather during the observing run. But note that "you can't tell the phase stability by looking out of the window"[4]—you must *observe* to find out how stable the phases are. It is difficult either to predict phase stability in advance (except to say that it is likely to be bad because a storm front is passing, for example) or to judge it during the observations by "looking out of the window". Clear blue skies do not guarantee stable phases. Thunderstorms do however guarantee unstable phases!

If your program has frequency agility, it is a good idea to monitor the phase of a strong calibrator for a few minutes on a long baseline as you start observing. Fluctuations of order a radian on a time scale of minutes are unmitigated bad news, and the only possible strategy is to move the observations to lower frequencies if this makes astronomical sense. The converse is not true, however. Short-term (minute-by-minute) phase stability to within a few degrees does not guarantee that the observations will be of good quality for synthesis. Synthesis imaging requires stability over the time scale of your calibration cycle (unless you are going to self-calibrate). You should therefore pay attention to the phase changes between *adjacent* scans of your calibrator, as well as to that within the scans, to assess whether you have the stability needed for synthesis. If the longer-term stability is marginal, i.e., of order 30°–40° between calibrations, you might consider using a faster calibration cycle.

Acknowledgments

I thank Ron Ekers, Rick Perley, Fred Schwab, Jacqueline van Gorkom, Arnold Rots, and many students who attended the 1982 and 1985 NRAO Summer Schools in Socorro, for valuable comments on earlier versions of this lecture.

[4] Words of wisdom widely promulgated during the early days of the VLA, and attributed to Barry Clark.

Appendix A: Considerations Specific to the VLA

To avoid burdening the main text with VLA-specific details that may obstruct the view of the lecture's general principles, I have collected as many of these details as possible into this Appendix. Readers who are interested in using other radio synthesis arrays may also find this Appendix helpful as a "worked example" of points discussed earlier.

A1. Choosing a VLA Configuration

a) Standard configurations.

An image made from untapered uniformly-weighted ≥ 4 hour tracks in a standard VLA configuration when foreshortening of the array is unimportant has a synthesized beam B with a half-power beamwidth given approximately by

$$\theta_{\mathrm{HPBW}} = 1''\!.25 \times \frac{1480}{\nu_0} \times 3.285^{n-1}, \tag{24–A1}$$

where ν_0 is the observing frequency in MHz and $n = 1, 2, 3$, or 4 for the **A**, **B**, **C**, or **D** configuration, respectively.

It is most important to choose a suitable combination of ν_0 and n when planning VLA observations. For example, consider a smooth two-dimensional emission region $30''$ across with a peak apparent brightness $I\Omega_s$ of 1 mJy per beam area on an untapered 20 cm image made with the VLA **B** configuration (resolution $\approx 4''\!.2$). The same region will have a peak apparent brightness of only 0.093 mJy per beam area on an untapered 20 cm image made with the same hour-angle coverage and u-v weighting in the **A** configuration (resolution $\approx 1''\!.3$). It could be detected at the 10σ level in about 16 min of integration at 50 MHz bandwidth in the **B** configuration (using the sensitivity data given in Table 24–2), but a 10σ detection in the **A** configuration using the same bandwidth would require about 31 hours of on-source integration! It is is therefore *extremely* important not to use a VLA configuration that gives a smaller beam area Ω_s than is strictly necessary, when studying extended emission.

Note also that the effects of spectral index and resolution combine to make extended *steep-spectrum* emission much harder to detect with a given VLA configuration at the higher frequencies. For example, suppose that an extended emission region has a peak intensity of 1 mJy per 'CLEAN' beam area in the VLA's **A** configuration at 20 cm—a 10σ detection would be made in 16 minutes at 20 cm. If the region has a ν^{-1} spectrum, the peak intensity in the **A** configuration at 6 cm would be 0.027 mJy per 'CLEAN' beam area and a 10σ detection at this frequency would require 160 hours of integration. The choice of observing frequency is therefore critical when trying to detect steep-spectrum extended emission using a given VLA configuration.

For sources with compact flat-spectrum components *and* extended steep-spectrum emission, the dynamic range needed to image the extended structure increases rapidly with increasing frequency. Suppose that the extended emission referred to in the previous example surrounded a 5 mJy point source with a ν^0 spectrum. The dynamic range required for 10σ detection of the extended structure would be $50:1$ in the **A** configuration at 20 cm. This is easy to obtain. The dynamic range required in the **A** configuration at 6 cm would be $1850:1$, a non-trivial goal without self-calibration.

b) Combinations of standard configurations.

Options other than the standard (**A, B, C, D**) configurations of the VLA are available. You will need to combine observations made in more than one VLA configuration if you require a *range* of baselines that exceeds the range provided by a standard configuration.

As the ratio of the longest to the shortest baseline in a standard configuration of the VLA is about $40:1$, each standard configuration can be used to image reliably up to $\theta_{\mathrm{LAS}} \approx 40\theta_{\mathrm{HPBW}}$ where θ_{HPBW} is given by Equation 24–A1 at the specified frequency. If the values of θ_{LAS} and θ_{HPBW} needed for your experiment do not *both* fall between θ_{HPBW} and $40\theta_{\mathrm{HPBW}}$ calculated from Equation 24–A1 for a given standard configuration and frequency, consider using more than one VLA configuration—unless you can extend the projected baseline range by observing at extreme hour-angles (where the u-v baselines may be significantly foreshortened). Observations requiring $\theta_{\mathrm{LAS}}/\theta_{\mathrm{HPBW}} > 40:1$ generally require more than one configuration, but some with $\theta_{\mathrm{LAS}}/\theta_{\mathrm{HPBW}} < 40:1$ have the same need—for example, if the optimum θ_{HPBW} falls mid-way between two resolutions given by allowed values of n and ν_0 in Equation 24–A1.

For example, Figure 24–3 shows the u-v coverage of the VLA at $+60°$ declination for 12 hours observing in the **A** configuration, and for 6 hours of **A** configuration observing combined with 6 hours in the **C** configuration. The "hole" at the center of the u-v coverage in Figure 24–3 could be filled by mixing data from the **A** and more compact configurations. Consider mixing standard-configuration observations for any source for which $\theta_{\mathrm{LAS}}/\theta_{\mathrm{HPBW}}$ is significantly greater than $40:1$. The total integration times needed in the different configurations should be computed separately; most projects will not need as long a total integration time in the more compact configurations as in the more scattered ones.

c) Hybrid configurations.

"Hybrid" configurations are available during reconfigurations, when the arms of the VLA may either be of different length, or have a non-standard assortment of long and short baselines. Some hybrid configurations provide wider ranges of u-v spacing than can a standard configuration (giving sensitivity to a wider range of angular scales). Some can help self-calibration of data from a compact configuration by providing unusually long spacings.

Hybrid configurations with long North arms are now regularly scheduled at the VLA. They are useful when imaging regions south of $\delta \approx -15°$, where the north–south extent of the u-v coverage of the standard configurations is seriously foreshortened by projection. Figure 24–4 shows the u-v coverage for the **B** configuration at $\delta = -40°$, compared with that of a hybrid configuration whose East and West arms are in the **B** configuration but whose North arm is in the **A** configuration. The spacings involving the longer North arm fill in a region around the v-axis that is left empty by the standard **B** configuration. This **A/B** hybrid is available briefly about every sixteen months, during a reconfiguration from **A** to **B**. Similar **B/C** and **C/D** hybrids are scheduled between the appropriate reconfigurations.

Perley (1981b) examined whether other hybrid VLA configurations could usefully extend the ratio of maximum to minimum baselines in synoptic observations with the VLA. In general, you get better u-v coverage by mixing data from two different standard configurations than by spending the same total time in *any* hybrid configuration. No other hybrid VLA configurations are regularly scheduled.

d) Sub-arrays.

"Sub-arrays" are nonstandard configurations obtained by dividing the VLA into smaller arrays that are devoted to different observing programs at the same time. The use of sub-arrays is generally not as efficient as time-sharing the entire VLA. The number of interferometer pairs in a sub-array is $N(N-1)/2$ where N is the number of antennas in the sub-array. Sub-arrays with 13 and 14 antennas therefore have 78 and 91 interferometers respectively, whereas a 27-antenna standard configuration has 351. An hour of observing in which two such sub-arrays each perform different tasks therefore produces 169 interferometer-hours of data. In contrast, two half-hours of observing, with the full VLA devoted to each task in turn, produce 351 interferometer-hours of data. Dedicating two roughly equal sub-arrays to different tasks thus reduces the amount of information gathered by a factor of about two, compared with time-sharing the whole VLA between the two tasks. This loss of information manifests itself as poorer sensitivity and u-v sampling in the sub-array data. The use of sub-arrays is therefore generally undesirable unless you need *strictly* simultaneous observations of strong sources at several frequencies (e.g., to measure instantaneous spectra of rapid variables) or to study many compact sources quickly with only modest demands on sensitivity and dynamic range (e.g., for astrometry of strong sources).

e) "Scaled" configurations.

Returning to Equation 24–A1, note that the scaling factor between "adjacent" VLA configurations (e.g., **B** and **C**) is 3.285. This factor is close to the ratios between the default VLA frequencies at 20 cm and 6 cm and between those at 6 cm and 2 cm. The VLA therefore

Figure 24–3. u-v coverage for $\delta = +60^\circ$ in the **A** configuration (12-hour tracks).

Figure 24–4. *u-v* coverage at $\delta = -40°$ with **(a)** the VLA East and West arms in **B** configuration and the North arm in **A** configuration, and **(b)** the entire VLA in **B** configuration.

has similar resolutions at 20 cm in the **A** configuration, at 6 cm in the **B** configuration, and at 2 cm in the **C** configuration. (Such rough three-frequency scalings also apply for the **B**, **C**, and **D** configurations, of course.) These scalings make the VLA a powerful tool for studies of the frequency-dependence of the properties of extended emission.

A2. Frequency Selection

Table 24–1 lists the frequency bands now available at the VLA.

Table 24–1.
VLA Observing Frequencies

Frequency Range	Band Name	System Temperature	Antenna Efficiency
0.3 to 0.35 GHz	P Band (90 cm)	125	0.40
1.26 to 1.73 GHz	L Band (20 cm)	60	0.51
4.5 to 5.0 GHz	C Band (6 cm)	50	0.65
8.0 to 8.8 GHz	X Band (3.6 cm)	30	0.58
14.4 to 15.4 GHz	U Band (2 cm)	116	0.52
22.0 to 24.0 GHz	K Band (1.3 cm)	150	0.43

The VLA continuum system lets you observe at two independent sky frequencies within each "band". This capability can be used to increase sensitivity, to fill in the u-v plane more densely by crude "bandwidth synthesis" (see Lecture 12, Section 1.1) or to study spectral or Faraday depth changes in your source across a "band" (the latter being especially worthwhile across the 'L' Band).

Interference is rarely detected or suspected above 1.8 GHz at the VLA. It is an important factor in choosing a continuum observing frequency within the L Band (1260 to 1730 MHz), particularly when using non-standard frequencies (e.g., when seeking to observe at the opposite edges of the band to determine Faraday rotation parameters). As spectral line observers do not have freedom to choose center frequencies for their projects, L Band interference may determine whether a given spectral line experiment is possible. There is self-generated interference throughout L Band at the VLA, mainly at the harmonics of 50 MHz; this internal interference should be below the noise in any continuum image made with an IF bandwidth $\Delta\nu > 6.25$ MHz, but can be a serious problem for spectroscopy.

Before using a non-standard frequency below 2 GHz, consult with VLA scientific staff (particularly Pat Crane, the VLA frequency coordinator) for advice and lore based on recent observers' experiences.

A3. Bandwidth

Users of extremely narrow bandwidths should be aware that when observing in continuum mode the VLA bandwidths narrower than 6.25 MHz suffer large closure errors because the quadrature networks do not work well. If such narrow bandwidths are essential for your observations, consider observing with the spectral-line system, where these problems are avoided. The VLA spectral-line system is not a panacea for eliminating chromatic aberration, however. It cannot (yet) support continuum polarimetry fully (and—because of the need to supply 'channel 0' for spectroscopy—it will not provide more than 3 "IFs" at 50 MHz total bandwidth even when completed in late 1989).

Chromatic aberration effects at the VLA are well estimated by the expressions given in Lecture 13 for a square bandpass and Gaussian tapering in the u-v plane. In the regime ($0 < \beta \leq 1$) where the amplitude reduction for a point source is moderate, the exact expression for the reduction (Lecture 13, Equation 13–23) can be approximated by

$$\frac{I}{I_0} \approx 1 - \frac{1}{3}\left(\frac{\gamma\Delta\nu\theta}{2\nu_0\theta_{\mathrm{HPBW}}}\right)^2 + \frac{1}{10}\left(\frac{\gamma\Delta\nu\theta}{2\nu_0\theta_{\mathrm{HPBW}}}\right)^4 , \tag{24–A2}$$

using the standard expansion for the error function, erf, at small values of its argument.

A4. AVERAGING TIME

The on-line visibility averaging time τ_a at the VLA may be set to any multiple of $1\frac{2}{3}$ sec, but the even multiples are preferable, for reasons associated with the VLA phase switching system. In practice, averaging for only 1.67 sec may give large ($\sim$ 10 mJy) correlator offsets, so the shortest τ_a recommended for all applications is 3.3 sec. The "default" values of the on-line visibility averaging time τ_a at the VLA are 10 sec for the **A** and **B** configurations and 30 sec for the **C** and **D** configurations.

The effects of the on-line averaging are permanent, because the averaged data are written to the archive tape created by the on-line system. It is therefore important to choose an on-line averaging time that limits time-averaging distortions to acceptable values, and is shorter than the expected atmospheric coherence time (provided the hardware and computer restrictions permit such a choice—spectroscopists must note Table 24–6). Longer averaging times may be used off-line, e.g., once the data have been calibrated.

For Gaussian tapering in the u-v plane, the intensity reduction due to a finite averaging time τ_a is given by Equation 13–39 with $\alpha = \gamma^2/\pi^2$, which we may write as

$$\frac{I}{I_0} \approx 1 - \frac{\gamma^2}{12}\omega_e^2\tau_a^2\left(\frac{\theta}{\theta_{\mathrm{HPBW}}}\right)^2 . \tag{24–A3}$$

Spectroscopists should use Equation 24–A3, with values of I/I_0 that are acceptable for their experiment, to determine the maximum acceptable averaging time.

Continuum observers may wish instead to use an averaging time $\tau_{\Delta\nu}$ that produces the *same* intensity reduction for a source near the pole as does an IF bandwidth $\Delta\nu$. This can be approximated (for small intensity reductions) by equating the terms in $\theta^2/\theta^2_{\mathrm{HPBW}}$ from Equations 24–A2 and 24–A3, obtaining

$$\tau_{\Delta\nu} \approx \frac{\Delta\nu}{\omega_e\nu_0} = 1.375 \times 10^4 \, \frac{\Delta\nu}{\nu_0} \text{ sec} . \tag{24–A4}$$

Equation 24–A4 gives a reasonable criterion for the maximum averaging time τ_a that should be used with an IF bandwidth $\Delta\nu$ for continuum observations at observing frequency ν_0.[5] Notice that $\tau_{\Delta\nu}$ does not depend on VLA configuration or on θ_{max}, owing to the first-order similarities between the bandwidth and time-average smearing for a polar source. Lecture 2 gave an intuitive basis for Equation 24–A4: for a polar source, time averaging smears the visibilities over an *angle* $\omega_e\tau_a$ in the u-v plane so its effects scale as $\sqrt{l^2+m^2}\,(\omega_e\tau_a)$, whereas those of bandwidth smearing scale as $\sqrt{l^2+m^2}\,(\Delta\nu/\nu_0)$.

You may have to exceed the value of $\tau_{\Delta\nu}$ calculated from Equations 24–A3 or 24–A4 because the VLA's on-line system cannot provide averaging times less than 1.67 seconds in any of the normal observing modes. Note also that the VLA on-line system averages all baselines with the same τ_a, and that the 'FILLER' program used to convey data to the DEC-10 off-line system imposes the *same* averaging time for the target and calibrator observations.[6] If the calibrator observations are only a few minutes each (as is often the case at low frequencies), averaging times longer than 1 min are undesirable because they permit only crude editing of the calibrator data. τ_a is also limited in many cases by the size of the final data set, especially for multiconfiguration continuum syntheses, or multichannel—continuum or spectral line—observing (Table 24–6). If storing the data set requires much more than the average disk space available per user in the off-line computers, data processing (and/or life with one's colleagues) becomes difficult. Averaging times τ_a are sometimes set longer than would otherwise be desirable, just to avoid "excessive" data volumes. Remember that one 512-byte "block" of disk in most computers that run AIPS stores 7.53 VLA visibility records, so a VLA continuum data set with both the AC and BD IFs uses about $93t_{\mathrm{int}}/\tau_a$ such disk blocks.

---

[5] It differs slightly from the formula in earlier versions of this lecture, due to my use here of more accurate expressions for the smearing effects from Lecture 13.

[6] This restriction has been removed in the 15OCT88 release of the corresponding AIPS task, 'FILLM'.

Most VLA continuum observing is done with 10 sec $< \tau_a <$ 1 min because of these constraints. For line work, the minimum averaging time and the number of spectral line channels are linked by computer usage rules at the VLA, but observers who will process their data away from the VLA may exceed these limits (if they can accommodate the resulting data sets at home!). Table 24–6 gives the rules that are in force in late 1988. Confirm them when planning a spectral line proposal—the details change with time, and changes are published in the *NRAO Newsletter* and the *VLA Observational Status Report*.

A5. The Primary Beam

To first order, the primary beam of the VLA antennas can be taken to be a Gaussian with FWHM equal to $90\lambda_{\rm cm}$ arcsec, and this approximation suffices for most proposal planning. In standard VLA software, the primary beam correction ($1/\mathcal{A}$, Lecture 3) is modeled by a power law expansion of the form $\frac{1}{\mathcal{A}} = \sum a_i X^{2i}$ where $X = \theta_{\rm arcmin}\nu_{\rm GHz}$ and the coefficients are $a_0 = 9.920378 \times 10^{-1}$, $a_1 = 9.956885 \times 10^{-4}$, $a_2 = 3.814573 \times 10^{-6}$, $a_3 = -5.311695 \times 10^{-10}$, and $a_4 = 3.9809630 \times 10^{-12}$ (Napier and Rots 1982). This correction is usually not applied at small values of X for which the Napier–Rots expansion gives $\mathcal{A} > 1$. It is not well determined at large values of X for which $\mathcal{A} < 0.15$.

A6. Sensitivity

Table 24–2 gives the theoretical r.m.s. noise on I and P images made at the VLA without tapering, using 27 antennas and the maximum interference-free continuum bandwidths, for integration times typical of "snapshots" (see below) and of more complete syntheses.[7] (Interference normally restricts observations at 92 cm to a 3 MHz bandwidth.)

Table 24–2.
Theoretical R.m.s. Noise on Images Made with 27 VLA Antennas*

Band Designation:	92 cm	20 cm	6 cm	3.6 cm	2 cm	1.3 cm**
	P	L	C	X	U	K
ΔS (1 sec, 1 MHz), Jy	1.87	0.60	0.45	0.31	1.14	2.40
Bandwidth $\Delta\nu$, MHz	3	50	50	50	50	50
R.m.s. noise in 5-min "snapshot", mJy/beam	2.4	0.19	0.15	0.10	0.37	0.78
R.m.s. noise in 12-hr integration, mJy/beam	0.20	0.016	0.012	0.008	0.030	0.064

*For two IFs and natural weighting. For uniform weighting, multiply all entries by 1.5 (for a first approximation).
**1.3 cm values are for the new (1988) receivers.

A7. "Snapshot" Mode

The "Y" layout of the VLA produces an *instantaneous* synthesized beam with a respectable shape and sidelobe level. It is therefore possible to do interesting science with brief observations if the targets are both bright and compact. Snapshot mode observing may be ideal for observers who wish to study statistical properties of large samples of sources (and also to overdose on synthesis image processing!).

In what follows, I consider a single "snapshot" to be an observation of $<$ 5 minutes' duration. Snapshots of $<$ 1 minute duration may even be appropriate if you want to image many ($>$ 1000) fields that are near to one another on the sky (so that antenna drive times are also short) and it does not matter if the occasional observation is abbreviated or even lost.

[7] Note that some system parameters (e.g., sensitivities) that affect these decisions will improve with time as a result of hardware upgrades, etc. NRAO publishes a *VLA Observational Status Report* that summarizes relevant system parameters at least once per year. You should check the most recent copy of this *Report* when planning a VLA proposal.

a) Limitations of "snapshot" mode.

The clearest limitation of snapshot observing is sensitivity (see Table 24–2); snapshot mode is suitable only for bright sources. Below 2 GHz, the high sidelobe levels of snapshot beams exacerbate the problems created by confusing sources, and snapshots of fields near the galactic plane using the more compact VLA configurations are frequently dominated by sidelobe clutter from confusing sources rather than by the noise that is quantified in Table 24–2. These problems are less severe above 2 GHz because of the smaller primary beam and the typical source spectrum (see Section A9 on "Confusion" below).

The second limitation of snapshot observing is the restricted angular size scale $\theta_{\rm LAS}$ over which the u-v coverage of a snapshot (e.g., Fig. 24–5) satisfies the sampling theorem and thus permits reconstruction of the correct sky brightness distribution. Table 24–3 estimates this limitation for snapshots on the meridian in the standard VLA configurations and frequencies, averaged over all declinations.

Table 24–3.
Approximate Values of $\theta_{\rm LAS}$ for a Single Snapshot*

	A	**B**	**C**	**D**
92 cm	$140''$	$8'$	$23'$	$50'$
20 cm	$30''$	$100''$	$5'$	$12'$
6 cm	$9''$	$30''$	$90''$	$200''$
3.6 cm	$6''$	$18''$	$53''$	$120''$
2 cm	$3''$	$10''$	$30''$	$70''$
1.3 cm	$2''$	$7''$	$20''$	$45''$

*Larger structures can be imaged by combining a few snapshots taken at different hour angles.

Polarization calibration may be difficult for short snapshot programs; it is not easy to verify the instrumental polarization calibration for a program whose total observing time is only a few hours, as this calibration requires at least three observations of a calibrator spanning a change in parallactic angle χ of $\Delta\chi \geq 90°$ (see Lecture 5, Section 7.1). "Standard" instrumental polarization parameters may then have to be used—note that these are available only for a few "standard" combinations of VLA observing frequencies and bandwidths (the default frequencies for 50 MHz bandwidths at 20 cm, 6 cm and 2 cm, and the default frequencies for 25 MHz and 12.5 MHz bandwidths at 6 cm). Position angle calibration may also be difficult if the standard polarization calibrators are not readily observable during the time allocated to a snapshot program. Snapshooters interested in polarimetry should ensure that suitable polarization calibration is possible when designing their program, by giving attention to its LST range and the choice of observing frequencies and bandwidths.

Snapshots are most effective when all sources are observed within about 2 hours of the meridian. At larger hour-angles, foreshortening of the array leads to poorer sampling of the u-v plane, elliptical synthesized beams, etc.

The time taken to calibrate a snapshot data set is determined mainly by the total observing time. Snapshot programs require the same calibration effort as simple synthesis programs of the same total duration. The image construction, deconvolution and display steps of snapshot observing can require large amounts of computer time and your time, however. As a snapshot image of a given target may be as large as a full synthesis image of the same field, snapshot programs also make heavy demands on disk storage. This can be especially true for snapshots made in the more compact configurations below 2 GHz, which are particularly prone to degradation by sidelobe clutter from confusing sources. Snapshooters must therefore be prepared to coordinate their data reduction requirements with those of other users, and to adopt efficient reduction strategies, including backing up of inactive images, beams and u-v data sets whenever possible.

b) Multiple snapshots versus extended snapshots.

Snapshots require phase stability only for the duration of the individual snapshot. There is therefore some debate over whether (for example) an observation requiring 15 minutes of

Figure 24–5. The u-v plane coverage for an instantaneous sampling of data for a source at $\delta = 30°$ and $H = 0$ by the 27-antenna VLA.

integration time $t_{\rm int}$ is better made as one continuous 15-minute observation or by combining the data from three separate 5-minute snapshots. A single 15-minute observation may give better dynamic range, because ionospheric or tropospheric phase gradients in the form of "wedges" may calibrate out of a single short observation, leaving only a position shift. In contrast, three shorter observations that are more dispersed in time and calibrated on a long calibrator cycle may encounter different wedges and so combine to give an image with poorer final dynamic range, unless self-calibration (Lecture 9) can be used. There are therefore some circumstances in which a single observation is preferable, as well as being easier to schedule.

The advantages of combining data from several shorter snapshots are (a) greater protection against total loss of the data for a given source through equipment failures or short-term bad weather, and (b) more even sampling of the u-v plane than in a single extended snapshot. Multiple snapshots are particularly useful when observing at wavelengths of 18 cm and longer in the **C** and **D** configurations, as they allow better imaging of confusing sources that may otherwise limit the achieved dynamic range (see Section 6 below). The single extended snapshot may be better for observations that must be made at low elevations, where phase "wedges" are more likely to arise, and if self-calibration cannot be used. This may be particularly true for observations of weak or complex low-declination sources for which the total hour-angle coverage is anyway limited by the short time that a given source is above the horizon.

A8. Spectral Line Observing

The VLA correlator has enough chips to produce 16 complex frequency channels at the nominal "50 MHz" (46 MHz) bandwidth. Halving the total bandwidth allows the chips to be "time-shared" so that the number of channels is doubled. This total bandwidth halving and time-sharing can be continued to a maximum of 512 channels, at which point further halvings of the total bandwidth leave the number of channels at a maximum of 512. Table 24–4 summarizes the resulting channel number and channel width options for a single IF, with and without on-line Hanning smoothing. Hanning smoothing to reduce the Gibbs phenomenon is strongly recommended for all observations that require high spectral dynamic range; it may be applied either on-line or off-line.

Table 24–4.
Total Channels in VLA Spectral Line Modes

	Without Hanning		With Hanning	
Bandwidth (MHz)	Channels	Separation (kHz)	Channels	Separation (kHz)
0.195313	512	0.381	256	0.763
0.78125	512	1.526	256	3.052
1.5625	512	3.052	256	6.104
3.125	256	12.207	128	24.414
6.25	128	48.828	64	97.656
12.5	64	195.313	32	390.625
25	32	781.25	16	1562.5
46	16	3125	8	6250

The above total numbers of channels may also be split evenly among two, or four, IFs. For example, if an observation is set up at 12.5 MHz total bandwidth with 2 IFs and no on-line Hanning smoothing, there will be 32 channels in each of the two spectra (i.e., 32 per spectrum for a total of 64 as shown in the table). The splitting among the IF channels is selected on the //DS input "card" by specifying the correlator "mode". Note that the first channel ("channel 0") of each IF contains the complex average of the central 75% of the band that the correlator has available (even if not all of that band has been selected).

The on-line system update of August 2, 1988 made available three new spectral line modes that support two IFs—modes 2AC, 2AD and 2CD. Other modes are planned that will use the entire correlator (including full polarization) but are not supported yet. All old modes (2A, 2C, 2D, 4A, 4D) that use only a fraction of the correlator are still supported at the time of writing this lecture, but will be phased out once all the new modes are in place—probably in the last quarter of 1989. Table 24–5 shows the VLA spectral line correlator modes that will be supported once the on-line system is fully upgraded, under the following headings:

Mode: Correlator Mode as expressed on the //DS input "card".

Supported?: "Yes", if it is available in the current (late 1988) system; "No" if it is not likely to be available until late 1989.

Number of IFs: The number and VLA designation of the IFs recorded on the archive tape.

For further details of correlator options and strategy decisions that involve them, read Rots (1988) or Sowinski and Braun (1988).

Table 24–6 gives the current averaging time restrictions that are imposed at the VLA to limit the burden on the off-line data reduction systems. These restrictions do not apply to data sets that will be reduced entirely outside the VLA off-line computers. These rules may change with time, and VLA staff and the *NRAO Newsletter* should be consulted for updates.

Table 24–5.
VLA Spectral Line Correlator Modes

Mode	Supported?	Number of IFs
1A	Yes	1 A
1B	Yes	1 B
1C	Yes	1 C
1D	Yes	1 D
2AB	No	2 A,B
2AC	Yes	2 A,C
2AD	Yes	2 A,D
2BC	No	2 B,C
2BD	No	2 B,D
2CD	Yes	2 C,D
4	No	4 A,B,C,D
4PA	No	4 A,C
4PB	No	4 B,D

Table 24–6.
VLA Averaging Time Restrictions

Channels	Minimum (sec)
1	3.3
8	5
16	7
32	15
64	60
128	120
256	240
512	480

A9. Confusion

The number of extragalactic sources N per square arc minute of sky with flux densities greater than S mJy at 6 cm can be written approximately as

$$N(>S) = 0.032 S^{-1.13} \tag{24–A5}$$

over the flux density range that is relevant for confusion calculations at the VLA (e.g., Ledden *et al.* 1980). The corresponding expression at 20 cm is

$$N(>S) = 0.10 S^{-0.9} \,. \tag{24–A6}$$

The analogs of these expressions for 2 cm and 1.3 cm are not known directly from measured source counts. They could be *estimated* from the 6 cm count in Equation 24–A5 by scaling flux densities to 6 cm with an effective mean spectrum of $\sim \nu^{-0.6}$.

Images made at 20 cm will contain, on average, one extragalactic source of flux density 110 mJy closer to the field center than the $15'$ HWHM of the primary beam. The 6 cm primary beam ($4\overset{'}{.}5$ HWHM) will contain, on average, one extragalactic source of flux density 2 mJy, the 2 cm beam ($1\overset{'}{.}85$ HWHM) a source of < 0.1 mJy and the 1.3 cm beam ($1'$ HWHM) a source of < 0.01 mJy. At 92 cm, there is always at least one several-Jy source in the primary beam and there may be many. Individual pathological cases aside, confusion is unlikely to be a problem at wavelengths shorter than 6 cm. At 20 cm and 6 cm it may be a problem in the VLA's more compact configurations. At 92 cm it is a problem in *all* configurations.

Confusion is particularly troublesome for snapshooters using the compact configurations below 2 GHz, because the sidelobes resulting from the "snowflake" u-v coverage of snapshots

(Fig. 24–8) are extensive. Images at low galactic latitudes are also more confused than those at high galactic latitudes. Snapshooters should plan to reduce their data using the ungridded subtraction algorithm ('MX' in AIPS) both because it permits imaging of multiple subfields and because it eliminates the effects of sidelobe aliasing.

A10. Calibration Details

a) Instrumental and atmospheric monitoring.

Reference sources should generally be chosen from the *VLA Calibrator List* maintained at the site by the NRAO staff, unless you are sure that a putative calibrator is unresolved in the VLA configuration that will be used, and has a position measured in the VLA reference system to better than 0.1 arcsec.

The length of time spent on each calibration scan should be enough to achieve a signal-to-noise (over the 26 baselines contributing to each antenna gain solution) commensurate with the required calibration accuracy. Never plan to calibrate for less than 1 minute at a time, however, as shorter calibrator scans may be lost as a result of unusually long settle-down times, etc. Significant atmospheric amplitude and phase fluctuations can occur on time scales of minutes, even at wavelengths of 6 cm and longer. When the Sun is active, *ionospheric* fluctuations will dominate at 18 cm and longer—they can also be rapid on long VLA baselines. It is completely impractical to adopt a *calibrator/target/calibrator* cycle that will guarantee following the fastest fluctuations of either kind. Keep in mind that *no* external referencing, no matter how rapid, can be *guaranteed* to remove atmospheric fluctuations from the data, and that time spent driving to and observing calibrators is time deleted from integration, and *u-v* coverage, on your target. You must decide for yourself how to play this particular roulette game during a given run.

Typical VLA observing programs spend from 5% to 10% of their time on calibration at the lower frequencies; more calibration may be needed at the higher frequencies where the calibrators are weaker and therefore need to be observed for longer total integration times. Calibration every 20 minutes or so will often follow the longer-term atmospheric fluctuations at 20 cm and 6 cm, especially in the more compact VLA configurations. Calibration every 10 minutes or so is safer at 2 cm and 1.3 cm, especially if the external calibrator is within a degree or so of the target. It should not be necessary to calibrate instrumental effects more rapidly than every 30 minutes at 20 cm or 6 cm.

Finally, note the significance of the choice of the gain table interval for the VLA off-line data base created by the 'FILLER' program *if you will not self-calibrate* your data. The off-line gain table interval (which you specify to the array operator at the time of the observations) sets the minimum time scale of instrumental or atmospheric fluctuations that can be corrected by an external calibration. (Self-calibration algorithms construct their own gain tables based on the integration time τ_{sc} that you specify for the gain determination.) The VLA default gain table interval of 10 minutes is adequate for calibration under a stable atmosphere, but shorter intervals are often more appropriate if you will rely on external calibration.

b) Flux-density calibration.

If the LST range of your observing run permits, observe 3C 286 for a few minutes at each observing frequency used for the targets, as 3C 286 is the flux-density standard to which all VLA measurements are ultimately referred. Failing this, observe 3C 48 (or consult with VLA staff about recent determinations of the amplitude gains of the antennas from other observations before finalizing your observing program). Do not simply take the most recent flux density for an arbitrary calibrator from the *VLA Calibrator List*, as many small-diameter sources are variable. The flux densities recorded in the *VLA Calibrator List* will rarely have been measured recently enough to determine the absolute flux-density scale for your observations; use them only to estimate the integration times needed to achieve the desired gain accuracy from your calibrator scans.

c) Polarization calibration.

To calibrate the *instrumental polarizations*, observe one unresolved calibrator, whether polarized or not, at least three times. These observations should be distributed so they cover a range in parallactic angle χ of $\Delta\chi \geq 90°$, to separate any polarization of the calibrator from the required instrumental terms (see Lecture 5). Programs involving long ($\geq$ 4 hr) syntheses of single sources will normally be able to derive the instrumental polarization calibration from the observations of the external gain calibrator. When determining the integration time for the instrumental polarization calibration, bear in mind that the leakage terms (the D's of Lecture 5) whose relative amplitudes and phases are to be determined will normally produce polarized intensities that are only a few percent of the flux density of the calibrator. The instrumental polarization calibration should be done at each frequency for which polarimetry is required. The most efficient way to do this is to cycle through the frequencies used for the target observations each time the array is pointed at the calibrator.

If the instrumental polarization calibration is omitted (e.g., because the observing session is too short, or the instrument misbehaved), you may be able to make the instrumental polarization corrections using the "standard" files of the necessary parameters that are maintained by the VLA staff. Note however that these are available only for a few combinations of observing frequency and bandwidth. If you omit an instrumental polarization calibration, you will be less able to determine small degrees of polarization, or to deconvolve polarized extended structures properly.[8]

To calibrate the *polarization position angle* scale, observe 3C 286 or 3C 138 at least once during your observing run at each relevant frequency. You will determine the apparent position angles of the linear polarization of these sources after you have finished observing and after calibrating the total intensity data. The difference between the apparent and the nominal values of these position angles is corrected by adjusting the phase difference between the left and right circular polarizations, using a procedure that is described in detail in the *VLA Cookbook*. It is advisable to alert the array operator to the presence of this calibration in your program, so that the observations of 3C 286 or 3C 138 can be extended or rescheduled if necessary to avoid losing them to equipment failure. Note that this calibration is *essential* if you wish to use your polarization position angle data.

At wavelengths of 18 cm and longer, the position angle calibration may appear to be time variable because of fluctuations in the ionospheric Faraday rotation (Lecture 5). If you will use the polarization position angle information at long wavelengths, it is worth monitoring one polarized calibrator *in the same part of the sky as your target(s)* throughout your observing, to see if its apparent position angle changes significantly. If this further calibration shows that the ionospheric changes are less than about 20°, it will probably be satisfactory to interpolate the observed position angle changes as a function of time when adjusting the relative phase of the left and right circularly polarized channels. If larger changes are seen, it may be possible to compensate for them using an ionospheric model and measured critical frequencies (by running the VLA's 'FARAD' program once the relevant critical frequency data have been received at the VLA—often several months after the observing). Except when the rotation changes are small ($< 20°$), the success of this repair cannot be guaranteed, however. The observation of the polarized calibrator can be a "warning light" for the existence of ionospheric Faraday rotation problems, not necessarily a means for correcting them. Applying FARAD's corrections to the data on this calibrator will also check whether they are indeed improving the angle calibration. Ionospheric effects will normally be negligible above about 4 GHz.

d) Self-calibration criteria.

To decide whether or not you will be able to self-calibrate VLA data, use Equation 24–4 with $N = 27$, i.e.,

$$\phi_{\rm rms} \approx \frac{13\Delta S}{S_{\rm sc}\sqrt{\Delta\nu\tau_{\rm sc}}}\ \text{degrees}\,, \qquad (24\text{–A7})$$

to estimate the phase error in a self-calibration interval $\tau_{\rm sc}$.

[8] Antenna-to-antenna polarization differences distort the polarization images in ways that do not satisfy the convolution theorem.

Figure 24–6. A sample of Item 19 from the standard VLA proposal cover sheet.

e) Special problems of P Band.

Calibration at P Band is particularly troublesome because of confusion in the primary beam. Confusing sources contribute several Jy or more to every field of view, and the use of narrow bandwidths for interference rejection implies that these sources contribute significantly to virtually all baselines. External calibrators must therefore be chosen for their strength rather than by nearness to the interesting field of view. Calibrators with > 25 Jy are preferable. Unless the calibrator is > 5 Jy, you may as well attempt self-calibration using the brightest sources in the field. Self-calibration with the traditional algorithms may not work under severe ionospheric conditions, where the size of the isoplanatic patch may become smaller than that of the primary beam. Perley (1988) has listed possible P Band calibrators.

Because ionospheric Faraday rotation at P Band is large and variable, polarization calibration is extremely difficult and cannot be guaranteed unless the ionosphere is exceptionally "quiet" during an observing run.

A11. Observing Proposals

All VLA observing proposals are refereed outside the NRAO before being considered by the scheduling committee. To receive high priority, the scientific goals of the project must favorably impress the external referees. A "highly-placed source who wishes to remain anonymous" notes that more concisely-written proposals are more likely to be refereed favorably, all else being equal.

Before you begin writing a proposal, it is worth checking whether any of your target sources have already been observed at the VLA. The NRAO routinely archives the raw (uncalibrated) visibility data from all VLA observations. General access to these data is restricted for 18 months after the end of the associated observing program. Thereafter, the raw data are available to any scientist who adequately justifies their use. Calibrated and processed data remain the property of the original observer and are not automatically indexed or archived at the VLA. Full catalogs of the observed sources, including relevant instrumental parameters, can be obtained from Teresa McBride at the VLA, or by accessing the DEC-10 disk files ACCUM.SRT[13,542] (all observations > 1 hour) or yyyy.SRT[13,542] (all observations in year *yyyy*) on the VLA DEC–10 computer, or from the PC-based program 'VLASORS' (see below). If you decide that the major parameters of an earlier observation are also appropriate for your work, you may either contact the original observers to use their calibrated data, or request use of the uncalibrated data from the VLA archive. Watch the *NRAO Newsletter* for possible updates about the archive and associated software.

The proposal cover sheet should be filled out in as much detail as possible. Filling out item 19 on the cover sheet (Fig. 24–6) fully for each target, or for typical targets, will lead you to consider the issues discussed in this lecture. Your entries here should show the referees and the scheduling committee that the proposal is well suited to the requested configuration(s).

When writing the proposal, plan your calibration strategy in detail. The time you request should include all the on-source integration time that you need, plus time for any and all

Figure 24–7. A histogram of durations of projects scheduled for VLA observations for the first two months of 1984, when the array was in the **B** configuration.

calibrations that you will do. Depending on the project, calibration time may need to be from 10% to 50% of the total request.

Figure 24–7 shows a typical histogram of the distribution of observing time allotted to successful VLA proposals. It covers the first two months of 1984, when the VLA was in the **B** configuration. The median observing time scheduled is 7 hours, reflecting how many proposals need less than full hour-angle tracks. Note however that some of the projects scheduled used more than 16 hours of observing time—well-justified long projects can successfully compete for time! The average number of proposals submitted to the VLA during 1983–87 was 466 per year, of which an average of 84 per year were rejected altogether. The average proposal scheduled received 55% of the requested observing time.

Proposals should be sent to the NRAO Director in Charlottesville well before the deadline for the desired configuration(s). The deadlines and the VLA configuration schedule are published regularly in the *NRAO Newsletter* and in the *AAS Newsletter*. Proposals may be submitted between the deadlines, and indeed the NRAO encourages this for several reasons—(a) the pressure of proposals for a given configuration influences the length of time that the VLA is scheduled to spend in that configuration, (b) early submission may let you reply to unfavorable referees' comments before the scheduling committee assigns time for the requested configuration(s), and (c) observers who submit early reduce the strain on the proposal processing system near the deadline.

A12. PC Software

Two PC-based software packages that are useful when planning VLA observing can be obtained on request from Sandra Montoya at the NRAO in Socorro.

a) VLASORS.

This self-documenting database query program will search either for calibrators in the *VLA Calibrator List* or for sources from the observation archive list. The *Calibrator List* may be searched by source name, frequency, flux density, or area of sky. The archive list may be searched by source name, frequency, position, proposer or project number. 'VLASORS' can be run on any IBM PC or compatible computer with a hard disk. The archive of all observations with total integration times longer than one hour requires 2.6 Mbytes of hard disk storage. The file of all observations requires 5 Mbytes. 'VLASORS' is distributed on three (> 60 min archive) or five (all observations) floppy diskettes.

b) VLAPLAN and VLAUVPL.

'VLAPLAN' (Bridle 1988) is a self-documenting, menu-driven spreadsheet that can be used to design continuum, spectral line, or bandwidth synthesis observing strategies for the VLA using the approaches and formulae described in this lecture. It organizes the relevant calculations into 'screens' that display the results of different parameter choices, offer advice on how to meet your imaging specifications within present VLA hardware and software limitations, and warn of parameter conflicts. 'VLAPLAN' also plots graphs of the bandwidth and time average smearing effects, of the primary beam correction, and of Gaussian source visibilities, scaled to the context of your parameter choices. It also warns of conflicts between the selected observing bandwidth and known interfering signals at L Band. 'VLAPLAN' requires an IBM PC or compatible with at least 320k of free RAM and a graphics board, and Lotus *1-2-3* (Version 2) or Borland *Quattro*.

'VLAPLAN' recommends VLA configuration(s) for your observing. It chooses the most compact configuration using only the observing frequency and the largest angular size and declination of your source. To save computation time, this recommendation is based on a table of the shortest projected baselines for *snapshots on the meridian* in each of the the four standard configurations, *at 10° intervals* in Declination. 'VLAUVPL' is an ancillary Lotus *1-2-3*, or *Quattro*, spreadsheet that computes the shortest projected baseline for *full synthesis* observations with the VLA at *any* Declination, but takes longer to run. It also generates u-v plots and graphs of $\sqrt{u^2 + v^2}$ versus hour-angle over the full range of hour-angles consistent with your prescribed elevation limits, to help you plan short-baseline coverage in detail.

References

Adams, D. (1979), *The Hitch-Hiker's Guide to the Galaxy*, Pan Books, London.

Bridle, A. H. (1988), "VLAPLAN—VLA observing strategy planner", VLA Computer Memorandum No. 179, NRAO.

Clark, B. G. (1981), "Orders of magnitude of some instrumental effects", VLA Scientific Memorandum No. 137, NRAO.

Ledden, J. E., Broderick, J. J., Condon, J. J., and Brown, R. L. (1980), "A confusion-limited extragalactic source survey at 4.755 GHz. I. Source list and areal distributions", *Astron. J.*, **85**, 780–788.

Napier, P. J. and Rots, A. H. (1982), "VLA primary beam parameters", VLA Test Memorandum No. 134, NRAO.

Perley, R. A. (1981b), "VLA hybrid configuration u-v plane coverage", VLA Scientific Memorandum No. 139, NRAO.

Perley, R. A. (1988), "Calibration of P Band data", VLA Scientific Memorandum No. 159, NRAO.

Rots, A. H. (1988), *A Short Guide for VLA Spectral Line Observers (Edition 7.0)*, NRAO.

Sowinski, K. and Braun, R. (1988), "Interim spectral line support", VLA memorandum dated August 12, 1988, NRAO.

Thompson, A. R. (1982a), "The response of a radio-astronomy synthesis array to interfering signals", *IEEE Trans. Antennas Propagat.*, **AP-30**, 450–456.

25. THE DESIGN OF APERTURE SYNTHESIS ARRAYS

R. M. Hjellming

1. Introduction

The goal of array design for aperture synthesis in radio astronomy is to obtain the best radio imaging instrument possible for the desired range of wavelengths and resolutions, for the lowest cost of construction and operation. While this general principle is obvious, the subject is complicated for a number of reasons. The first reason is that "best" is both difficult to define and, like "beauty", very dependent upon the "eye of the beholder". The second reason is that the answers depend upon the types of structure present in the radio sources that one wishes to image, hence they are dependent upon the types of science for which the array is to be optimized. Finally, in general you get what you pay for, so every array design is limited by real or assumed constraints for construction and operational budgets.

The purpose of this lecture is to discuss many of the theoretical and practical factors that are involved in array design. Much of the material and many of the ideas have been developed in the context of the work on the Millimeter Array which is currently planned as NRAO's next major instrumental project after the VLBA. After discussing the performance factors and the practical considerations that are important, I will consider the quantitative formulation of array performance factors. I will then describe the major types of ideal arrays for both one-dimensional and two-dimensional antenna distributions. The problems involved in trade-offs between performance and array design by minimizing costs will also be discussed. Finally, I will attempt to summarize the major designs that one would choose for a few specific circumstances. However, it is more important to understand the reasons for the various optimal choices than it is to decide which is the "perfect" array design.

2. The Parameters of Aperture Synthesis Arrays

Consider the design of an array with N antennas, where each antenna has a diameter D. Assume that the array is, or can be, instrumented for operation at wavelengths λ, where $\lambda_{\rm min} \leq \lambda \leq \lambda_{\rm max}$. For present purposes I define the goal of array design to be the determination of N, D, and the spatial distribution of the N antennas in one or more configurations. I will denote the maximum antenna separation in any single configuration as B, but there may be a range of B's used at different times if a very wide range of resolutions is desired.

The antenna diameter determines the size of the angular field that can be imaged with a single pointing of the array antennas. I will discuss this field of view in terms of the antenna half-power beamwidth, $\theta_{\rm ant} \approx \lambda/D$. The maximum resolution of an array of maximum size B is given by $\theta_{\rm beam} \approx \lambda/B$. While the smallest angular size scales that can be "seen" in the image are of the order of

$\theta_{\rm beam}/2$, and the locations of smaller sources can be determined with greater positional accuracy, I will use $\theta_{\rm beam}$ to describe the maximum resolution of an array. The same array is capable of imaging with $\theta > \theta_{\rm beam}$ by selectively weighting the data from the inner over the outer parts of the array. Reasonable imaging requires at least two image pixels per synthesized beam,[1] so the "size" of the image for a single pointing of the array is $\sim (2B/D)^2$ pixels (more if there are significant sources outside the half-power point of the antenna beams).

Figures 25–1a and 25–1b show the schematic relationship between the distribution of antennas inside an area of diameter B and the corresponding aperture plane (u-v plane) obtained when data from all of the N antennas are cross-correlated to obtain $N(N-1)/2$ instantaneous antenna pairs or baselines. These figures also illustrate three important aspects of aperture synthesis that are essentially independent of array design:

(1) Each u-v data point must be represented twice, with each visibility measurement at (u, v) being matched with its complex conjugate at $(-u, -v)$, so that the image computed from the data has real intensities;

(2) There are no data points corresponding to antenna pairs with separations greater than B; and

(3) The smallest possible separation of antenna pairs for a given array is D, corresponding to some antenna pairs with centers separated by one antenna diameter.

Figures 25–1c and 25–1d illustrate a specific application for a "filled circle" array of 21 antennas that I will discuss later in the lecture, with 25–1c showing the antenna distribution and 25–1d showing the appropriate gridding in a 17×17 array which is 60% "sampled" with a single brief observation.

Figure 25–1 does not show three of the important parameters of array design: (i) the number of occupied "cells" in the aperture plane, $N_{\rm occ} \leq \pi(B/D)^2$; (ii) the distribution or number density of data points in the aperture plane ($N(u, v) = N(q)$ where $q = \sqrt{u^2 + v^2}$); and (iii) the mean ($\langle n \rangle_{\rm m}$) or harmonic mean ($\langle n \rangle_{\rm hm}$) number of data points in each "cell" in the aperture plane; however, Figure 25–1d shows the very high proportion of occupied cells that occurs for "packed" arrays.

The square "gridding" of Figure 25–1a indicates that an ideal aperture of size B would be synthesized if antennas could be placed at all possible locations separated by D. As I will show, this extreme does indeed approach the case of a single large antenna of diameter B, except for the inevitable hole of size $\geq D$ in the center of the aperture plane. This ideal limit is achieved only by using $N_{\rm max} \approx \pi(B/D)^2$ antennas, and only the smallest of possible configurations can achieve this ideal. One might describe one goal of array design as trying to obtain the best result possible with fewer antennas. One indication that $N_{\rm max}$ antennas are not required is the fact that while an instantaneous, "filled" aperture plane would require $N_{\rm max}$ antennas, a smaller number of antennas would

[1] I.e., two points across any one-dimensional (or linear) cut through the central lobe of the beam. — *Eds.*

Figure 25–1. Schematic diagrams illustrating: **(a)** the distribution of antennas in an array; **(b)** the corresponding aperture or u-v plane for the pairs formed by these antennas; **(c)** a specific "packed" array of 21 antennas; and **(d)** the associated snapshot sampling of that 21-antenna array.

accomplish the same purpose by making continuous measurements over a long period of time so that the rotation of the Earth would cause a smaller number of antenna pairs to rotate into all the ideal locations of a filled array. Finally, the number of required antennas can be reduced even further by matching the capabilities of each array to only some of the subsets of problems to be solved. I therefore now turn the problem around and ask about the capabilities of specific array configurations. However, before doing this, let me discuss some of the performance factors of arrays. Some of these are quantitative in nature, but (unfortunately) some important ones have not yet been quantified in any simple way.

3. Performance Factors for Aperture Synthesis Arrays

The previous section, which emphasized the generalities of array design, directly or indirectly specified some of the major performance factors for arrays. Table 25–1 introduces most of the performance factors that I will discuss.

Table 25–1. Array Performance Factors and Their Determining Parameters

Array Performance Factor	Determining Parameters
Sidelobe Level Distribution	Minimization of the number of "holes" in the u-v plane; uniformity of hole distribution.
Range of Angular Resolutions	Maximum and minimum antenna separations, wavelengths instrumented, and range of configuration sizes (B's).
Sensitivity to point sources	Total collecting area $= \pi N D^2/4$, antenna aperture efficiency, correlator quantization efficiency, bandwidth, system temperature of electronics, and observing time.
Surface-brightness sensitivity	Intrinsic array resolution, distribution of the number of data points in the aperture plane, $N(u,v) = N(q)$ where $q = \sqrt{u^2+v^2}$, and point source sensitivity.
Optimization for mosaicing images larger than antenna beam	Maximize ND rather than ND^2.
Beam sidelobe levels	Distribution of missing data in aperture plane.
Optimization for deconvolution	Distribution of missing data and signal-to-noise noise distribution in u-v plane.
Optimization for self-calibration techniques and/or mathematical simplicity of data sampling/processing	Most of the above, plus optimization for redundancy or non-redundancy in the u-v plane.
Optimize for combination with non-aperture synthesis images	Minimum spacings matched to single antenna imaging, viewed in this lecture as a design problem for single-antenna imaging.

Previous lectures in this Summer School have covered many of the performance factors listed in Table 25–1. One more category could be listed: optimization for geometric simplicity. Because of practical aspects, like cost, this is a general characteristic of arrays that are cheaper to implement. Some of the factors listed in Table 25–1 are so important for array design that is worthwhile to summarize the ones that can be quantified.

3.1. Holes in the u-v plane.

Define $N_{\rm occ}$ to be the number of occupied "cells" in a u-v plane gridded to satisfy the criterion of two points or pixels per synthesized beam. The maximum possible value of this parameter for an array is $N_{\rm occ,max} = (\pi/4)(2B/D)^2$, so the fraction of occupied cells is $f_{\rm occ} = N_{\rm occ}/N_{\rm occ,max} \leq 1$.

3.2. Point source sensitivity.

If we adopt the notation

$$\begin{aligned}
k_B &= \text{Boltzmann's constant,} \\
T_{\rm sys} &= \text{system temperature,} \\
\eta_a &= \text{antenna aperture efficiency,} \\
\eta_q &= \text{correlator quantization efficiency } (= 0.82 \text{ for a three-level correlator}), \\
\Delta\nu &= \text{signal bandwidth, and} \\
\Delta t &= \text{integration time (total observing time for a single pointing position),}
\end{aligned}$$

then (from Lecture 7, Eq. 7–33) a single antenna pair has an r.m.s. point source sensitivity of

$$\sigma_c = \frac{4\sqrt{2}\, k_B T_{\rm sys}}{\eta_a \eta_q \pi D^2 \sqrt{\Delta\nu\, \Delta t}}\,, \tag{25–1}$$

so that, for N antennas and N_p combined polarizations, the r.m.s. point source sensitivity (sensitivity to weak point sources, that is) in an image is

$$\sigma = \frac{\sigma_c}{\gamma \sqrt{N_p N(N-1)/2}}\,, \tag{25–2}$$

where $\gamma \equiv \sqrt{\langle n\rangle_{\rm c}/\langle n\rangle_{\rm m}}$ is a factor that depends upon the distribution of the data points in the aperture plane, with $\langle n\rangle_{\rm c}$ being the "average" number of data points in each occupied "cell" and $\langle n\rangle_{\rm m}$ being the mean number of data points in each occupied cell. The meaning of the abovementioned "average" is complicated. If there is no tapering in the aperture plane during the process of computing an image, then $\langle n\rangle_{\rm c} = \langle n\rangle_{\rm m} = \left(\sum n_j\right)/N_{\rm occ}$, the arithmetic mean of the number of data points in occupied cells for the case of "natural" weighting of data, and $\langle n\rangle_{\rm c} = \langle n\rangle_{\rm hm} = 1/\left(N_{\rm occ}\sum 1/n_j\right)$, the harmonic mean number of data points in occupied cells for the case of "uniform" weighting. These two types of means are the simplest parametrization of various kinds of distributions of data in the aperture plane. Thus one finds that three of the important parameters of array design are $N_{\rm occ}/N_{\rm max}$, $\langle n\rangle_{\rm m}$, and $\langle n\rangle_{\rm hm}$ (or γ).

3.3. Surface-brightness sensitivity.

Surface-brightness sensitivity for an array is much more complicated because it depends upon the synthesized beam, which itself depends not only upon the distribution of data in the aperture plane—but also upon the weighting and tapering applied to the aperture plane when the image is computed from the data. One defines surface-brightness sensitivity in terms of the r.m.s. brightness temperature, which is given by

$$\Delta T_b = \frac{\sigma/(2k_B)}{\lambda^2 \Omega_{\rm beam}} = \frac{\sigma B^2}{2k_B \Gamma}\,, \tag{25–3}$$

where $\Omega_{\rm beam}$ is the solid angle of the synthesized beam integrated out to the first null and $\Gamma = \Omega_{\rm beam}/(\lambda/B)^2$ is another geometric correction factor that is a ratio of the synthesized beam solid angle to a measure of the smallest possible synthesized beam solid angle. The value of Γ changes arbitrarily for tapered images but has a well-defined mean for images made with "natural" or "uniform" weighting. The values of γ and Γ for these cases are array design parameters.

3.4. Minimum observing time per array pointing.

The speed with which an array achieves desired goals is clearly one of the factors to be considered in array design. The time it takes for Earth rotation to "fill" the aperture plane is an obvious parameter; it may take anywhere from 12 hours for linear arrays, to very short observing times for two-dimensional arrays with excellent instantaneous sampling of the aperture plane. Another fundamental time scale for an array is the time it takes to achieve a desired r.m.s. noise level in the image plane. If $\sigma_{(\rm mJy)}$ is the desired level of point source sensitivity, then from Equation 25–2 one can show that the time per array pointing (in minutes) needed to achieve this level is

$$\Delta t_{(\rm min)} = \frac{2.92\langle T_{\rm sys}\rangle^2}{\sigma^2_{(\rm mJy)}\gamma^2 D^4_{(\rm m)}\Delta\nu_{(\rm GHz)} N_p N(N-1)/2}\,, \tag{25–4}$$

where the bandwidth is in GHz, the antenna diameter is in meters, and the system temperature is in degrees Kelvin. Equations 25–2 and 25–4 argue for the smallest values of $\langle T_{\rm sys}\rangle$, the average system temperature during the observations, and the largest possible values of γ, $\Delta\nu$, and the product ND^2.

3.5. Minimum observing time for a large source or field.

If one poses the image sensitivity problem in terms of the observing time Δt_s needed to attain the desired noise level in an image of a radio source (or field of observation) much larger than the antenna beam, then the answer changes. This is because the number of different pointings of the array, each with an integration time given by Equation 25–4, is itself given by

$$\Delta t_s = \frac{\Omega_m}{\Omega_{\rm beam}}\Delta t_{(\rm min)} \propto \frac{\langle T_{\rm sys}\rangle^2}{(\lambda\sigma_{(\rm mJy)}\gamma N D)^2 \Delta\nu\, N_p}\,, \tag{25–5}$$

where one assumes a large number of antennas and that Ω_m is the size of the field to be mosaiced. Figure 25–2 shows an example of a packed array of 21 antennas observing one of the 16 fields in a particular "mosaic". Unlike the case of a

Figure 25–2. Schematic diagrams illustrating a 21-antenna packed array observing one field of a 16-field mosaic.

single field, the size of the antenna beam affects the time needed to complete the observations for "mosaicing" an image, with the result that it depends upon the wavelength of observation; thus the product ND is the quantity to be maximized in the array design if only minimum noise level for mosaiced images is the prime goal of the array.

3.6. Beam sidelobe level.

Sidelobe levels of synthesized beams are major parameters of array design. Cornwell (1985) has shown that one can estimate the sidelobe level of the synthesized beam solely from a knowledge of $N(u,v)$ for a particular array and observing situation. One again gets different answers for the cases of "natural" and "uniform" weighting, and for different tapering of the aperture plane. The r.m.s. sidelobe level is estimated from

$$s_{\mathrm{na}}^2 = \frac{\sum_j \sum_k (T(u_j, v_k) N(u_j, v_k))^2}{\left(\sum_j \sum_k T(u_j, v_k) N(u_j, v_k)\right)^2}, \tag{25–6}$$

for the case of natural weighting, where the sums are taken over all occupied "cells" in the aperture plane. The r.m.s. sidelobe level for uniform weighting is obtained by replacing $N(u,v)$ in Equation 25–6 by 1 for occupied cells and zero for unoccupied cells. A low r.m.s. sidelobe level is desired for highest imaging

Figure 25–3. The r.m.s. sidelobe level is plotted as a function of the number of antennas N for a VLA-like three-armed radial configuration, with open circles for the uniform weighting case and filled circles for the natural weighting case (and normalization to a common point at $N = 27$).

quality, and the deconvolution algorithms that effectively remove sidelobes work best if the sidelobe levels are low.

The s_{na} and s_{un} computed from Equation 25–6 for $N(u, v) = N(u, v)$ or 1, respectively, are meaningful only for comparative purposes. Figure 25–3 shows a plot of these two parameters as a function of N, with normalization of unity at the 27-antenna case of the VLA.

3.7. Beams for different density distributions of data in the aperture plane.

The synthesized beam for any aperture synthesis array, for any particular observing scenario, is obtained from the two-dimensional Fourier transform of the aperture-plane data assuming each visibility measurement has the constant amplitude and zero phase of a strong, unresolved or point source at the phase reference or phase tracking position of the array. It is therefore straightforward to compute synthesized beams. This then allows evaluation of sidelobe levels on specific size scales. The other aspects of array design that can be determined from such calculations are the consequences of different distributions of data points in the aperture plane, a factor that most affects the range of surface-brightness sensitivities that can be obtained with a single configuration. The most important aspect of this is the relations between the radial distribution, $N(u, v) = N(q)$, and the resulting ideal beam shape. For spherically symmetric distributions the two-dimensional Fourier transform of the aperture plane is replaced by the Hankel transform, that is

$$I_{\mathrm{beam}}(\theta) = 2\pi \int_0^{q_{\mathrm{max}}} N(q) J_0(2\pi q\theta) q \, dq \,, \qquad (25\text{–}7)$$

where $q_{\max} = B/\lambda$ and J_0 is the first-kind Bessel function of 0 order. As I will show, there are a number of Hankel transform pairs that correspond to certain classes of arrays. Thus once one considers a particular array design, it is simple to determine its characteristic $N(q)$ distribution, which can often be approximated by a simple analytic function. One then has an analytic result for the Hankel transform that describes the inner beam shape that is important for measurement of surface brightness. For example, VLA-like Y's with similar antenna distributions along the arms have $N(q) \propto \exp(-q/q_{\max})$, which Hankel transforms to $I_{\text{beam}}(\theta) \propto \left(1 + (\theta/\theta_0)^2\right)^{-3/2}$ for the "natural" weighting case; and for circular arrays, or arrays with data processing with "uniform" weighting, the $N(q) =$ constant Hankel transforms to $I_{\text{beam}}(\theta) \propto J_1(2\pi\theta/\theta_0)/\theta$.

Figure 25–4 shows the $N(q)$ distributions for three cases as obtained from a 21-antenna simulation, with Figures 25–4a and 25–4b showing the two cases mentioned in the previous paragraph. All have the "hole" in the center that can be filled only with non-interferometric data. In the case of the randomized circle "only" the outer 96% of the angular extent of the u-v plane is "constant". Figure 25–4c shows the important case of a packed array, corresponding to the array in Figure 25–1c, where the triangular $N(q)$ distribution of a single antenna is similar except for the "hole" in the center. In all cases, larger numbers of antennas would make the empirical curves more closely approach the theoretical ones.

3.8. Self-calibration for redundant or non-redundant arrays.

Self-calibration techniques play a major role in improving both the quality of aperture synthesis data and the corresponding dynamic range of the resulting deconvolved images. Without delving into this complicated question in any detail, it is worth mentioning that there are two extreme views on one aspect of array design if optimized for self-calibration techniques. On one extreme all arrays should be designed with a large degree of redundancy—that is, duplication of measurements for the same "cells" in the aperture plane—while on the other extreme non-redundancy is to be emphasized. With a high degree of redundancy because antenna locations are in simple steps of a single minimum spacing (such as exists for the Westerbork Synthesis Radio Telescope and the compact portion of the Australia Telescope) one can solve for calibration errors by comparing measurements that should be identical. With minimum redundancy, a case much more closely approached by the VLA than the previously mentioned instruments, one has better instantaneous beams; in addition the type of self-calibration using intermediate models of the radio image, generally the "clean" components of a deconvolved image, works very well. Both techniques work well for instruments for which they are best suited. I will emphasize the primacy of non-redundant arrays, mainly because of a desire for excellent imaging/self-calibration for both short and long observations.

3.9. Optimum matching with single-antenna imaging information.

The problem of imaging with a combination of aperture synthesis data and data derived from single-antenna observations is very important; however, it is mainly a question of relative antenna size and relative sensitivity for the antenna

Figure 25–4. Examples of the theoretical and empirical distributions of $N(q)$ vs. q for 21-antenna arrays corresponding to: **(a)** a VLA-like three-armed radial distribution; **(b)** a circular array; and **(c)** a packed circle 90 m in diameter filled with 21 ten-meter antennas.

doing the single antenna observations. The conventional wisdom of recent years has been that the latter antenna should be 2–3 times the diameter of the antennas used for the aperture synthesis data; however, it has recently been argued that even antennas of the same size could be used to achieve the desired results. This question is very important for the design of future aperture synthesis arrays, since obtaining data for the "hole" in the aperture plane is very important when imaging extended radio sources.

3.10. Optimization for geometric simplicity.

An abstract concept, which has real value, is that of "geometric simplicity". The costs of land, transportation systems, and electronic communications systems are all reduced if antennas are distributed in simple ways, with minimum complexity in the connectivity between antennas.

4. Practical Factors in Array Design

The idealized methods of optimizing the factors discussed in the previous section may be not be practical. Thus the choice of the design of a particular aperture synthesis array will always entail compromises due to realistic considerations, such as limits on construction or operational budgets.

4.1. Antenna diameter and the number of antennas.

A commonly used rule-of-thumb is that the cost per antenna is proportional to D^α, where α is somewhere between two and three. Recent antenna construction projects suggest that use of this rule, with $\alpha \approx 2.7$, might not lead one too far astray for antennas in the 20- to 35-meter class. Outside of this range a different power-law exponent α should be used—or, better yet, a more sophisticated model that accounts for the different scaling laws applicable to the costs of the various constructional components (panels, different parts of the backup structure, and foundations, for example).[2] It would be very helpful if one could assume, with some confidence, the validity of a cost vs. diameter model before considering trade-offs between N and D. One needs also to be aware of the desirability of maximizing ND or ND^2, depending upon whether one will mainly be imaging sources much larger than the antenna beams, or whether one will largely be imaging sources with size scale less than or on the order of the antenna beam.

4.2. Minimizing the number of antenna stations.

The costs of foundations, supplying electrical power, and the necessary systems for communicating with antennas for both antenna/electronics control and the acquisition of data at a central computer site are, to first order, proportional to the number of stations from which antennas can be operated. This is one of the obvious cost factors that affect both the number of antennas operating

[2] Quoting from von Hoerner (1987): "When scaling a design to other sizes, different members scale with different exponents of the telescope diameter D. Omitting all details: the many small members keep the bar area fixed and just go with D. Longer ones without much load keep the slenderness fixed and go with D^2. All those which carry the main loads of survival winds (ice, snow) keep the stress fixed and go with D^3. If a few members must be beefed-up for observational winds, they keep the deformation fixed and go with D^4. Altogether, we found from some examples an exponent of about 2.6 for the whole telescope."

in any one configuration and the number of different configurations of antennas that are practicable.

4.3. Antenna transportation system.

Except for the case in which the array operates in only one configuration, changes in configuration size and/or type generally require moving some or all of the N antennas to different stations. For large and heavy antennas, such as those of the VLA, this led to a design involving a twin-track rail transportation system. Under these circumstances one wants to minimize the costs of roadbeds, rails, etc.—so there are strong reasons for a design like the VLA **Y**-configuration, which minimizes the scale of the transportation system for an two-dimensional array. These costs are minimized even more for one-dimensional arrays. For smaller and lighter antennas the transporters that carry/pull antennas between stations may need only a reasonable system of connecting roads. In all cases, the cost of the antenna transportation system must be included in evaluating the costs of constructing and operating a flexible array.

4.4. Topography and site characteristics.

It is obvious that the cost, or even the feasibility, of an antenna transportation system can be limited by the characteristics of potential site locations. Soil characteristics, the existence of ravines, and the cost of available space all affect the costs of the location of array(s). The VLA is located on a very large, relatively flat area with good, stable soil characteristics—a lake bed of relatively recent (but not too recent!) origin.

Arrays for millimeter and sub-millimeter operation require as high and dry sites as practical because of the effects of atmospheric transparency at these wavelengths; not only does one lose source intensity due to atmospheric absorption, but the emission of the roughly 300 K atmosphere becomes a significant factor in the system temperature—which becomes exponentially dependent upon the atmospheric transparency. The latter factor also creates a need for excellent beam/imaging characteristic during the short time periods, and small zenith angle ranges, when transparencies are lowest. On the other hand, ideal and accessible high mountains with large, flat plateaus at the top are not easy to find; however, they may be important enough to force compromises with ideal theoretical configuration. This point will come up again when I describe a few theoretically ideal configurations. Figure 25–11 compares a theoretically ideal five-armed radial configuration with an example of a multi-armed array placed upon a hypothetical high plateau.

4.5. Data processing costs.

Before the advent of extensive spectral line aperture synthesis and modern deconvolution and self-calibration techniques it was reasonable to regard the costs of computers and computer software as a small part of construction and operational budgets. However, the computing needs for the best aperture synthesis techniques have grown to be so large that these factors must be included in budgets.

4.6. Construction costs vs. operational costs.

Finally, it is obvious, and I have mentioned a number of times above, that the costs of an array are not only in construction, but also in the costs of operation. It is well known that maintenance of large numbers of complex electronics systems is very labor-intensive, requiring large manpower budgets. The remote sites that may be astronomically ideal can be very expensive to operate.

The "bottom line" is that the choice of an array design involves a large number of theoretical and practical trade-offs. In the end one is generally, on the one hand, dealing with a minimization of costs as a function of N for a given D^{α} dollars/antenna; and on the other hand, one is trying to maximize ND or ND^2, or some combination thereof.

5. Major Types of "Ideal" Array Designs

Let us consider the characteristics of five different cases that are "ideal" under certain circumstances:

(1) linear or one-dimensional arrays;
(2) VLA-like radial configurations with three (*cf.* Fig. 25–5a) or five arms (Fig. 25–11a);
(3) two-dimensional non-redundant arrays (*cf.* Fig. 25–5b);
(4) randomized circular arrays (*cf.* Fig. 25–5c); and
(5) packed circular arrays (*cf.* Fig. 25–1c).

5.1. Linear arrays.

Most of the early arrays used in aperture synthesis in radio astronomy were one-dimensional or linear in the sense that all antennas were located along a straight line. These arrays require long observations, ideally 12 hours in length, to sample the two dimensions of the aperture plane reasonably. Because antennas cannot observe radio sources below the horizon, at lower declinations (for higher latitudes) one could not obtain good sampling in the N–S direction in the aperture plane. Such designs are, however, an excellent way to obtain redundant or non-redundant aperture distributions with small numbers of antennas. This is the reason most pioneering aperture synthesis instruments, such as the Cambridge and Westerbork arrays, were built as linear arrays. This type of design was an excellent choice for the compact portion of the Australia Telescope because it is optimal for the extreme negative declinations for which the Northern Hemisphere arrays are least optimal. The **T**-array, consisting of a linear East–West distribution of antennas complemented by some antennas on a North–South line, is an obvious extrapolation of an E–W array with better u-v sampling in N–S directions; it begins to have some of the good characteristics of three-armed radial distributions of antennas.

5.2. Radial power-law distributions of antennas.

The VLA is an example of a three-armed radial configuration in which the antennas on each arm, for each configuration, are located at distances from the center which are proportional to $n^{1.728}$, where $n = 1, 2, \ldots, N/3$ is the order of antennas from the center. The aperture-plane distributions for arrays like this can be approximated by $N(q) \propto \exp(-q/q_{\max})$, leading to a core beam shape

Figure 25–5. Antenna distributions corresponding to: **(a)** a VLA-like three-armed radial distribution of 21 antennas; **(b)** a non-redundant 2-dimensional array with antennas at random locations, analogous to a "crystalline" array design; and **(c)** a circular array with antennas at random locations.

of $I_{\rm beam}(\theta) \propto \left(1+(\theta/\theta_0)^2\right)^{-3/2}$ for "natural" weighting—a beam which has a sharp core, Lorentzian-like wings, and a $\theta_{\rm HPBW}$ which is almost exactly 50% larger than it is for "uniform" weighting. The "uniform" weighting case imposes $N(q) \approx$ constant, and again according to Equation 25–7 the Hankel transform of this function gives a beam like $I_{\rm beam}(\theta) \propto J_1(2\pi\theta/\theta_0)/\theta$. Figure 25–5a shows a 21-antenna example of a VLA-like array with the same power-law distribution for each arm, and Figure 25–6 shows examples of aperture-plane sampling for a VLA-like **Y** of 21 antennas and the corresponding "natural" and "uniform" weight beams for two cases: a very short "snapshot" observation and $\pm$4-hour tracking. Note the larger sidelobes for the "snapshot" case and the associated broad base of the main beam, which reflects the "Lorentzian"-like behavior of the Hankel transform of the "natural" exponential $N(q)$ distribution. Even with uniform weighting, some of the "natural" distribution of the VLA-like sampling appears in the snapshot beam, and is replaced by the $J_1(x)/x$ beam characteristic only for a well-filled uniform or uniformly weighted aperture.

Figure 25–7 shows the horizon-to-horizon aperture-plane coverage of a 27-antenna, three-armed, VLA-like configuration at (surprise!) the latitude of the VLA. It is characteristic of three-armed radial configurations that the outer 6-fold asymmetries seen in the cases for $\delta \leq 40°$ will appear. This can be removed somewhat by adopting a five-armed radial configuration, which will fill in the gaps considerably because the analogous asymmetries are 10-fold. Finally, for lower declinations the horizon limitations result in considerable lack of sampling along North–South orientations. The only way to obtain symmetric aperture-plane coverage for these cases is to elongate the antenna distribution by a factor of the order of 2–3. This is done in the VLA case by having periods with hybrid arrays in which the North arm is in a configuration a factor of 3.28 larger than the other two arms.

5.3. Non-redundant two-dimensional arrays.

An interesting case with excellent aperture-plane coverage is a two-dimensional array in which the antennas are located at essentially random locations inside a particular area. The work on "crystalline" arrays by Cornwell showed that when one attempts to obtain an ideal "beam" from a 2–D distribution of a small number of antennas, one places antennas at locations that, to the eye, appear random. Because it is difficult to design "crystalline" arrays with large numbers of antennas, a simple random number generator in two-dimensions, subject to the constraint that antennas not overlap, was used to generate analogous configurations. Figure 25–5b shows a sample of such a 21-antenna array, and Figure 25–8 shows the snapshot and full coverage aperture planes and beams (for the same natural and uniform weight cases as shown in Fig. 25–7). One sees from Figure 25–8 that the uniform weight beams for snapshot and full-coverage cases are very similar. For the natural weight cases there are obviously worse sidelobes because of the obvious "holes" in the aperture plane; these holes are removed with more extended observing; however, it is obvious that the worst remaining holes can be removed with judicious re-location of a few antennas.

The obvious practical defect of the non-redundant two-dimensional config-

Figure 25–6. The aperture planes, naturally weighted beam profiles, and uniform weighted beam profiles for a 21-antenna, VLA-like radial three-armed power-law distribution shown for both snapshot observations of a source at 60° declination and ±4-hour tracking.

Figure 25–7. The aperture planes for a range of declinations for a 27-antenna, three-armed radial distribution of antennas corresponding to the case of the VLA at a latitude of 34°.

Figure 25–8. The aperture planes, naturally weighted beam profiles, and uniform weighted beam profiles for a 21-antenna, non-redundant two-dimensional antenna distribution and observations of a source at 60° declination, shown both for snapshot observations and ±4-hour tracking.

Figure 25–9. The aperture planes, naturally weighted beam profiles, and uniform weighted beam profiles for a 21-antenna randomized circle antenna distribution and observations of a source at 60° declination, shown both for snapshot observations and ±4-hour tracking.

Figure 25–10. The aperture planes, naturally weighted beam profiles, and uniform weighted beam profiles for a 21-antenna packed configuration (with $2B/D = 18$) and observations of a source at $60°$ declination, shown both for snapshot observations and ± 4-hour tracking.

urations is the extensive transportation, power, and data communications networks required for each configuration.

5.4. Circular arrays.

In many senses the ideal configuration is a circular array in which antennas are located at random locations (excluding overlapping antennas) on a circle. Figure 25–5c shows a 21-antenna configuration of this type, and Figure 25–9 shows the aperture planes and beams for snapshots and ±4-hour coverage. The main utility of this case is that all examples have basically the same main beam and that even the sidelobes of the worst case snapshots are "randomly" located. Unlike the non-redundant 2–D case just discussed, a power/communications system for a series of circular arrays is not overly complicated.

5.5. Packed arrays and near-equivalence to very large antennas.

In Figure 25–1c and 25–1d I gave an example of packing 21 antennas in a configuration with a value of $2B/D \approx 18$. The corresponding snapshot and ±4-hour aperture planes and beams are shown in Figure 25–10. As has already been mentioned, the $N(q)$ distribution of packed arrays of this type approach the characteristics of single antennas of size B, except for the omnipresent hole in the center, as shown in Figures 25–1 and 25–10 for the array in Figure 25–4c.

6. "Quality" Factors For Arrays

Determining the relative "goodness" of different array designs is, as has been discussed, partly qualitative and partly quantitative. Let me illustrate quantitative comparison of arrays by considering three compact arrays of 21 antennas that were part of the initial phase of the Millimeter Array Design Study (Hjellming 1988, Brown and Schwab 1988). Assuming a bandwidth of 1 GHz, an average system temperature of 200° K, and both 2-minute snapshots and 8-hour tracking about the zenith, Table 25–2 lists the major quantitative parameters of three different arrays for source observations at a declination of 60°. The first configuration is a 300-meter three-armed radial configuration (*cf.* Figs. 25–4a, 25–5a, and 25–6). The second is a 300-meter configuration where the antennas are at random locations on a circle (Figs. 25–4b, 25–5c, and 25–9). The third is a configuration of antennas at random locations inside a 90-meter circle (Figs. 25–1c, 25–1d, 25–2, and 25–10). Table 25–2 lists the quantitative parameters discussed in previous sections for these arrays.

The comparison of the three-armed radial configuration with the randomized circle would lead one to conclude that the randomized circle is significantly better. This is largely because the values of the geometric correction factors, γ and Γ, are close to unity for both uniform and natural weighting cases. The highest-resolution images and best sensitivity are obtained under these circumstances. On the other hand, if the primary goal is imaging with arrays of different size scales, then there are advantages in matching configurations that have a ratio of a factor of 40 from the inner part of the u-v plane to the outer part of the u-v plane. This once again illustrates that no array design is optimal for all purposes and that choices must be made by a compromise between practicality and the primary scientific goals.

Table 25–2.
Summary of Quantitative Array Parameters for Three Arrays With $N = 21$, $D = 10$ m.

Configuration	3-Armed Radial		Randomized Circle		Filled Circle	
B (meters)	300		300		90	
D (meters)	10		10		10	
$2B/D$	60		60		18	
Observing Time	8^h	2^m	8^h	2^m	8^h	2^m
N_{occ}	2804	362	3418	410	308	204
f_{occ}	0.708	0.091	0.863	0.104	0.890	0.589
$\langle n\rangle_m$	7.10	0.57	5.90	0.51	65.3	1.03
$\langle n\rangle_{hm}$	1.88	0.53	3.94	0.51	11.7	0.79
γ	0.51	0.98	0.82	1.00	0.42	0.88
Γ	5.8	2.1	1.1	1.0	1.9	1.3
s_{na}	0.0325	0.0565	0.0200	0.0499	0.0717	0.0798
s_{un}	0.0189	0.0526	0.0171	0.0494	0.0570	0.0700
$\theta_{beam,na}/\lambda_{(mm)}$	$1''\!.30$	$1''\!.20$	$0''\!.51$	$0''\!.49$	$2''\!.09$	$1''\!.99$
$\theta_{beam,un}/\lambda_{(mm)}$	$0''\!.54$	$0''\!.82$	$0''\!.48$	$0''\!.49$	$1''\!.53$	$1''\!.75$
s_{na} (mJy)	0.079	1.22	0.079	1.22	0.079	1.22
s_{un} (mJy)	0.154	1.28	0.096	1.23	0.187	1.40
$\Delta T_{b,na}$ (mK)	0.64	11.5	4.07	69.1	0.25	4.2
$\Delta T_{b,un}$ (mK)	7.08	25.8	5.62	68.8	0.46	6.2

Any array at intermediate geographic latitudes, with the same size scale N–S as E–W, will have the problem of synthesized beams that are elongated when observing at lower declinations. The main solution to this problem is the use of configurations that are elongated in the N–S direction. For an array at a latitude of the order of 30° this means a N–S elongation by about a factor of three. This is achieved with multi-armed radial arrays, like the VLA, by operating in hybrid arrays with some arms one size-scale of configuration elongated to the North or South compared with the other arms. For "circular" or packed arrays the only solution is use of arrays with a 3 : 1 ellipticity. Such arrays then require observing over a range of a few hours in hour-angle to get symmetric beams for higher-declination sources.

7. Conclusions

With the obvious presumption that optimization for particular types of scientific problems may be more important than general considerations, there are a few general conclusions that can be drawn about arrays designs.

If the dominant imaging problems involve mosaicing a large number of observing fields together, then one wants the antenna diameter D to be as small as possible.

If the number of antennas N is small it is almost obvious that East–West linear arrays, or **T**-shaped arrays, will be best. This probably occurs for $N \leq 10$. If transport between configurations of different size scales is important, then it is probably true that for $N \geq 15$ one should build arrays that are randomized circles or ellipses elongated N–S by a factor depending upon geographic latitude; however, if antenna transport is expensive, as it was for the VLA because of

Figure 25–11. The antenna distributions for: **(a)** a 21-antenna, VLA-like radial five-armed power-law distribution; and **(b)** a topography-limited array with very similar characteristics, shown on a hypothetical high plateau.

the size/weight of the antennas, then 3-armed or 5-armed radial configurations (Figs. 25–5a and 25–11a) are probably preferable.

Sometimes other factors can be more important. A particular example is aperture synthesis at millimeter wavelengths. The fact that the signal-to-noise ratio is proportional to system temperature, and that this ratio increases exponentially with the amount of precipitable water vapor above an antenna or array, means that it is very important to have as high-altitude a site as possible. Figure 25–11b shows a sketch of a high-altitude plateau for which the topography limits the possible larger configurations. In practice, there is little significant difference between the aperture synthesis properties of the arrays in Figures 25–11a and 25–11b.

I believe that in this lecture I have covered most of the factors that affect the choice of designs of aperture synthesis arrays. This is not a simple problem, and array builders always must make choices that represent compromises between their realistic and their more naive goals.

References

Brown, R. L. and Schwab, F. R., Eds., *The Millimeter Array Design Concept*, MMA Design Study Vol. II, NRAO, Charlottesville, January 1988.

Cornwell, T. J. (1985), "Quality indicators for the MM array", Millimeter Array Memorandum No. 18, NRAO.

Hjellming, R. M. (1988), "The concept of the NRAO millimeter array", in *Science with a Millimeter Array*, A. Wootten and F. R. Schwab, Eds., MMA Design Study Vol. I, NRAO, Charlottesville.

von Hoerner, S. (1987), "What next? Suggestions for future radio telescopes", NLSRT Memorandum No. 2, NRAO, Charlottesville; also see later memoranda in this series, which relates to the design of large, steerable radio telescopes.

REFERENCES

For the reader's convenience, all references cited in this volume are listed below. However, to find which references pertain to a specific topic one might prefer to consult the reference section at the end of each relevant lecture.

Copies of memoranda in the VLA and VLBA technical and scientific series are available on request from Sandra Montoya, NRAO, P.O. Box O, Socorro, NM 87801, U.S.A.

Adams, D. (1979), *The Hitch-Hiker's Guide to the Galaxy*, Pan Books, London.

Alef, W. and Porcas, R. W. (1986), "VLBI fringe-fitting with antenna-based residuals", *Astron. Astrophys.*, **168**, 365–368.

Andrews, H. C. and Hunt, B. R. (1977), *Digital Image Restoration*, Prentice–Hall (Englewood Cliffs, NJ).

Archer, J. W. (1979), "Modifications to the VLA front ends for solar observations", Addendum to VLA Electronics Memorandum No. 181, NRAO.

Armstrong, J. W., Coles, M., Kojima, M., and Rickett, B. J. (1986), "Solar wind observations near the Sun", in *The Sun and the Heliosphere in Three Dimensions*, R. G. Marsden, Ed., D. Reidel (Dordrecht, Holland), pp. 59–64.

Arsac, J. (1955), "Nouveau réseau pour l'observation radioastronomique de la brillance sur la Soleil à 9530 Mc/s", *C. R. Acad. Sci.*, **240**, 942–945.

Bagri, D. S. (1986), "Gain residuals in the VLA", Draft.

Begeman, K. G. (1988), *HI Rotation Curves of Spiral Galaxies*, Ph. D. Thesis, University of Groningen.

Berge, G. L. (1965), *An Interferometric Study of Jupiter's Decimeter Radio Emission*, Ph. D. Thesis, Caltech.

Bevington, P. R. (1969), *Data Reduction and Error Analysis for the Physical Sciences*, McGraw–Hill, New York.

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, 1982*, A. R. Thompson and L. R. D'Addario, Eds., NRAO, (Green Bank, WV).

Bignell, R. C. and Napier, P. J. (1978), "Preliminary polarization calibration of the VLA at 6 cm", VLA Test Memorandum No. 125, NRAO.

Booth, R. S., Kus, A. J., Norris, R. P., and Porter, N. D. (1981), "Observations of a circumstellar shell around the OH/IR star OH127.8−0.0", *Nature*, **290**, 382–384.

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

Bos, A. (1984), "On ghost source mechanisms in spectral line synthesis observations with digital spectrometers", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 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. (1981), "21-cm line studies of spiral galaxies. I. Observations of the galaxies NGC 5033, 3198, 5055, 2861 and 7731", *Astron. J.*, **86**, 1791–1824.

Bracewell, R. N. (1966), "Optimum spacings for radio telescopes with unfilled apertures", 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., and Thompson, A. R. (1973), "The Stanford five-element radio telescope", *Proc. IEEE*, **9**, 1249–1257.

Bracewell, R. N. and Roberts, J. A. (1954), "Aerial smoothing in radio astronomy", *Aust. J. Phys.*, **7**, 615–640.

Braun, R. (1988), "Mosaicing with high dynamic range", Millimeter Array Memorandum No. 46, NRAO.

Braun, R. and Walterbos, R. A. M. (1985), "A solution to the short spacing problem in radio interferometry", *Astron. Astrophys.*, **143**, 307–312.

Bregman, J. D. and Felli, M. (1976), "The use of the Westerbork Synthesis Radio Telescope for solar observations at 21 cm", *Astron. Astrophys.*, **46**, 41–48.

Bridle, A. H. (1988), "VLAPLAN—VLA observing strategy planner", VLA Computer Memorandum No. 179, NRAO.

Brosche, P., Wade, C. M., and Hjellming, R. M. (1973), "Precise positions of radio sources. IV. Improved solutions and error analysis for 59 sources", *Astrophys. J.*, **183**, 805–818.

Brown, R. L. and Schwab, F. R., Eds., *The Millimeter Array Design Concept*, MMA Design Study Vol. II, NRAO, Charlottesville, January 1988.

Carter, W. E., Robertson, D. S., and MacKay, J. R. (1985), "Geodetic radio interferometric surveying: application and results." *J. Geophys. Res.*, **90**, 4577–4587.

Chikada, Y., Ishiguro, M., Hirabayashi, H., Morimoto, M., Morita, K.-I., Miyazawa, K., Nagane, K., Murata, K., Tojo, A., Inoue, S., Kanzawa, T., and Iwashita, H. (1984), "A digital FFT spectro-correlator for radio astronomy", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 387–404.

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

Christiansen, W. N. and Warburton, J. A. (1955), "The distribution of radio brightness over the solar disk at a wavelength of 21 centimetres. III. The quiet Sun—two-dimensional observations", *Aust. J. Phys.*, **8**, 474–486.

Chu, T. S. and Turrin, R. H. (1973), "Depolarization properties of offset reflector antennas", *IEEE Trans. Ant. Propagat.*, **AP-21**, 339–345.

Clark, B. G. (1973a), "VLA telescope pointing analysis", VLA Computer Memorandum No. 104, NRAO.

Clark, B. G. (1973b), "Ephemeris routines for the VLA—specification considerations", VLA Computer Memorandum No. 105, NRAO.

Clark, B. G. (1973c), "Curvature of the sky", VLA Scientific Memorandum No. 107, NRAO.

Clark, B. G. (1976), "Gridding and signal-to-noise ratios", VLA Scientific Memorandum No. 124, NRAO.

Clark, B. G. (1978), "Closure—some examples", VLA Electronics Memorandum No. 171, NRAO.

Clark, B. G. (1980), "An efficient implementation of the algorithm 'CLEAN'", *Astron. Astrophys.*, **89**, 377–378.

Clark, B. G. (1981), "Orders of magnitude of some instrumental effects", VLA Scientific Memorandum No. 137, NRAO.

Clark, B. G. (1982a), "The new IAU precession procedures", VLA Scientific Memorandum No. 145, NRAO.

Clark, B. G. (1982b), "Large field mapping", 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).

Cooper, B. F. C. (1970), "Correlators with two-bit quantization", *Aust. J. Phys.*, **23**, 521–527.

Cornwell, T. J. (1980), *The Mapping of Radio Sources from Interferometer Data*, Ph. D. Thesis, University of Manchester.

Cornwell, T. J. (1981), "An error analysis of calibration", VLA Scientific Memorandum No. 135.

Cornwell, T. J. (1983), "A simple method of stabilizing the clean algorithm", *Astron. Astrophys.*, **121**, 281–285.

Cornwell, T. J. (1984a), "Is Jaynes' maximum entropy principle applicable to image reconstruction?", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 291–296.

Cornwell, T. J. (1984b), "Broadband mapping of sources with spatially varying spectral index", VLB Array Memorandum No. 324, NRAO.

Cornwell, T. J. (1984c), "Mosaicing with the mm array", Millimeter Array Memorandum No. 32, NRAO.

Cornwell, T. J. (1985), "Quality indicators for the MM array", Millimeter Array Memorandum No. 18, NRAO.

Cornwell, T. J. (1988), "Radio-interferometric imaging of very large objects", *Astron. Astrophys.*, **202**, 316–321.

Cornwell, T. J., Anantharamaiah, K. R., and Narayan, R. (1988a), "Spatial coherence in a scatter-broadened image—observations of 3C 279 close to the Sun", in *Radio Wave Scattering in the Interstellar Medium*, J. M. Cordes, B. J. Rickett and D. C. Backer, Eds., AIP Conference Proceedings **174**, American Institute of Physics, New York, pp. 92–96.

Cornwell, T. J., Anantharamaiah, K. R., and Narayan, R. (1988b), "The propagation of coherence in scattering—an experiment using inter-planetary scintillation", *J. Opt. Soc. Amer.*, submitted.

Cornwell, T. J. and Evans, K. F. (1985), "A simple maximum entropy deconvolution algorithm", *Astron. Astrophys.*, **143**, 77–83.

Cornwell, T. J. and Napier, P. J. (1986), "Self-correction of telescope surface errors using a correlating focal plane array", in *Radio Astronomy from Space*, K. W. Weiler, Ed., Workshop Vol. No. 18, NRAO (Green Bank, WV), pp. 215–219.

Cornwell, T. J. and Napier, P. J. (1988), "The focal plane coherence function of an imaging antenna and its use in measuring and correcting aberrations", *Radio Science*, **23**, 739–748.

Cornwell, T. J. and Wilkinson, P. N. (1981), "A new method for making radio maps with unstable radio interferometers", *Mon. Not. Roy. Astron. Soc.*, **196**, 1067–1086.

Cotton, W. D. (1979), "A method of measuring compact structure in radio sources using VLBI observations", *Astron. J.*, **84**, 1122–1128.

Cotton, W. D., Geldzahler, B. J., Marcaide, J. M., Shapiro, I. I., Sanromá, M., and Rius, A. (1984), "VLBI observations of the polarized radio emission from the quasar 3C 454.3", *Astrophys. J.*, **286**, 503–508.

Crane, P. C. (1982), "Signal analysis of a correlation interferometer", VLA Scientific Memorandum No. 140, NRAO.

Crochiere, R. E. and Rabiner, L. R. (1983), *Multirate Digital Signal Processing*, Chapter 4, Prentice–Hall (Englewood Cliffs, NJ).

Croom, D. L. (1971), "Solar microwave bursts as indicators of the occurrence of solar proton emission", *Solar Phys.*, **19**, 152–170.

D'Addario, L. R. (1979), "Observing the Sun with the VLA", VLA Scientific Memorandum No. 130, NRAO.

D'Addario, L. R. (1984), "Minimizing storage requirements for quantized noise", VLBA Memorandum No. 332, NRAO.

D'Addario, L. R., Thompson, A. R., Schwab, F. R., and Granlund, J. (1984), "Complex cross correlators with three level quantization: design tolerances", *Radio Science*, **19**, 931–945.

Davenport, W. B. and Root, R. L. (1958), *An Introduction to the Theory of Random Signals and Noise*, McGraw–Hill, New York.

Dewdney, P. (1987), "Investigating the use of tiltmeters to correct VLA antenna pointing—Part I. Initial measurements and analysis", VLA Test Memorandum No. 148, NRAO.

Dulk, G. A., Bastian, T. S., and Kane, S. R. (1986), "Two-frequency imaging of microwave impulsive flares near the solar limb", *Astrophys. J.*, **300**, 438–447.

Dym, H. and McKean, H. P. (1972), *Fourier Series and Integrals*, Academic Press, New York.

Dyson, F. J. (1975), "Photon noise and atmospheric noise in active optical systems", *J. Opt. Soc. Am.*, **65**, 551–558.

Ekers, R. D. (1984), "The almost serendipitous discovery of self-calibration", in *Serendipitous Discoveries in Radio Astronomy*, Proceedings of a Workshop Held at the NRAO on May 4, 5, 6, 1983, K. I. Kellermann and B. Sheets, Eds., NRAO (Green Bank, WV), pp. 154–159.

Ekers, R. D. and Little, L. T. (1971), "The motion of the solar wind close to the Sun", *Astron. Astrophys.*, **10**, 310–316.

Ekers, R. D. and Rots, A. H. (1979), "Short spacing synthesis from a primary beam scanned interferometer", Proc. IAU Coll. 49, *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., D. Reidel (Dordrecht, Holland), pp. 61–66.

Ekers, R. D. and van Gorkom, J. H. (1984), "Spectral line imaging with aperture synthesis radio telescopes", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 21–32.

Erickson, W. C., Mahoney, M. J., and Erb, K. (1982), "The Clark Lake Teepee-Tee telescope", *Astrophys. J. Suppl. Ser.*, **50**, 403–420.

Escoffier, R. P. (1979), "Correlator system observers manual", VLA Technical Report No. 39, NRAO.

Ewen, H. I. and Purcell, E. M. (1951), "Radiation from galactic hydrogen at 1,420 Mc./sec.", *Nature*, **168**, p. 356.

Fomalont, E. B. (1973), "Earth-rotation aperture synthesis", *Proc. IEEE*, **61**, 1211–1218.

Fomalont, E. B. and Sramek, R. A. (1977), "The deflection of radio waves by the Sun", *Comments on Astrophys.*, **7**, 19–33.

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

Frater, R. H. (1984), "The Australia Telescope—the radio astronomer's highway to the future", *Proc. Astron. Soc. Australia*, **5**, 440–445.

Frieden, B. R. (1972), "Restoring with maximum likelihood and maximum entropy", *J. Opt. Soc. Am.*, **62**, 511–518.

Gary, D. E. and Hurford, G. J. (1987), "Multifrequency observations of a solar active region during a partial eclipse", *Astrophys. J.*, **317**, 522–533.

Gary, D. E. and Zirin, H. (1988), "Microwave structure of the quiet Sun", *Astrophys. J.*, **329**, 991–1001.

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

Granlund, J., Thompson, A. R., and Clark, B. G. (1978), "An application of Walsh functions in radio astronomy instrumentation", *IEEE Trans. Electromag. Compat.*, **EMC-20**, 451–453.

Greisen, E. W. (1976), "On the effects of convolution in the *u*-*v* plane", VLA Scientific Memorandum No. 123, NRAO.

Greisen, E. W. (1979), "The effects of various convolving functions on aliasing and relative signal-to-noise ratios", VLA Scientific Memorandum No. 131, NRAO.

Greisen, E. W. (1983), "Nonlinear coordinate systems in AIPS", AIPS Memorandum No. 27, NRAO, Charlottesville.

Grossman, A. and Muhleman, D. O. (1988), in preparation; reported by J. Eberhart, *Science News*, **134** (1988) p. 170.

Gull, S. F. and Daniell, G. (1978), "Image reconstruction from noisy and incomplete data", *Nature*, **272**, 686–690.

Habbal, S. R. and Harvey, K. L. (1988), "Simultaneous observations of 20 centimeter bright points and He I 10830 Å dark points in the quiet Sun", *Astrophys. J.*, **326**, 988–996.

Hamaker, J. P., O'Sullivan, J. D., and Noordam, J. E. (1977), "Image sharpness, Fourier optics, and redundant spacing interferometry", *J. Opt. Soc. Am.*, **67**, 1122–1123.

Hanbury Brown, R. A. (1974), *The Intensity Interferometer*, Taylor and Francis, London.

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

Hjellming, R. M. (1988), "The concept of the NRAO millimeter array", in *Science with a Millimeter Array*, A. Wootten and F. R. Schwab, Eds., MMA Design Study Vol. I, NRAO, Charlottesville.

Hjellming, R. M. and Basart, J. P. (1982), "The theory of the instrument", Chapter 2 in *An Introduction to the NRAO Very Large Array*, R. M. Hjellming, Ed., NRAO.

Högbom, J. (1974), "Aperture synthesis with a non-regular distribution of interferometer baselines", *Astrophys. J. Suppl. Ser.*, **15**, 417–426.

Högbom, J. A. and Brouw, W. N. (1974), "The synthesis radio telescope at Westerbork. Principles of operation, performance and data reduction", *Astron. Astrophys.*, **33**, 289–301.

Hudson, J. A. (1977), "An analysis of aberrations of the VLA radio synthesis telescope", Chapter 5, Ph. D. Thesis, The American University, Washington, D.C.

Hudson, J., Plambeck, R., and Welch, W. J. (1987), "Aperture efficiency enhancement in a Cassegrain system by means of a dielectric lens", *Radio Science*, **22**, 1091–1101.

Jaynes, E. T. (1982), "The rationale of maximum entropy methods", *Proc. IEEE*, **70**, 939–952.

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), "A phase sensitive interferometer technique for the measurement of the Fourier transforms of spatial brightness distributions of small angular extent", *Mon. Not. Roy. Astron. Soc.*, **118**, 276–284.

Kattenberg, A. and Palagi, F. (1983), "One-dimensional high time resolution solar observations with the Westerbork Synthesis Radio Telescope", *Astron. Astrophys.*, **125**, 1–5.

Kellermann, K. I. and Thompson, A. R. (1985), "The Very Long Baseline Array", *Science*, **229**, 123–130.

Killeen, N. E. B., Bicknell, G. V., and Ekers, R. D. (1986), "The radio galaxy IC 4296 (PKS 1333−33) I: Multi-frequency VLA observations", *Astrophys. J.*, **302**, 306–336.

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

Kundu, M. R. (1965), *Solar Radio Astronomy*, John Wiley & Sons, New York.

Kundu, M. R., Rao, A. P., Erskine, F. T., and Bregman, J. D. (1979), "High-resolution observations of the quiet Sun at 6 centimeters using the Westerbork Synthesis Radio Telescope", *Astrophys. J.*, **234**, 1122–1136.

Ledden, J. E., Broderick, J. J., Condon, J. J., and Brown, R. L. (1980), "A confusion-limited extragalactic source survey at 4.755 GHz. I. Source list and areal distributions", *Astron. J.*, **85**, 780–788.

Linfield, R. P. (1986), "A modified hybrid mapping technique for VLBI data", *Astron. J.*, **92**, 213–218.

Love, A. W., Ed. (1978), *Reflector Antennas*, IEEE Press, New York.

Mayer, C. E., Davis, J. H., Peters, W. L., and Vogel, W. J. (1983), "A holographic surface measurement of the Texas 3.9-m antenna at 86 GHz", *IEEE Trans. Instrum. Meas.*, **IM-32**, 102–109.

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

Michelson, A. A. (1890), "On the application of interference methods to astronomical measurements", *Phil. Mag.*, Ser. 5, **30**, 1–21.

Middleton, D. (1960), *An Introduction to Statistical Communication Theory*, McGraw–Hill, New York.

Moffet, A. T. (1968), "Minimum redundancy linear arrays", *IEEE Trans. Antennas Propagat.*, **AP–16**, 172–175.

Moran, J. M. (1973), "Spectral line analysis of very-long-baseline interferometric data", *Proc. IEEE*, **61**, 1236–1242.

Moran, J. M. (1976), "Very long baseline interferometric observations and data reduction", in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, pp. 228–260.

Muller, R. A. and Buffington, A. (1974), "Real-time correction of atmospherically degraded telescope images through image sharpening", *J. Opt. Soc. Am.*, **64**, 1200–1210.

Napier, P. J. and Rots, A. H. (1982), "VLA primary beam parameters", VLA Test Memorandum No. 134, NRAO.

Napier, P. J., Thompson, A. R., and Ekers, R. D. (1983), "The Very Large Array: Design and performance of a modern synthesis radio telescope", *Proc. IEEE*, **71**, 1295–1320.

Narayan, R. and Nityananda, R. (1984), "Maximum entropy—flexibility versus fundamentalism", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 281–290.

Narayan, R. and Nityananda, R. (1986), "Maximum entropy image restoration in astronomy", *Ann. Rev. Astron. Astrophys.*, **24**, 127–170.

Pearson, T. J. and Readhead, A. C. S. (1984), "Image-formation by self-calibration in radio astronomy", *Ann. Rev. Astron. Astrophys.*, **22**, 97–130.

Perley, R. A. (1981a), "The effect of bandwidth on the synthesized beam", VLA Scientific Memorandum No. 138, NRAO.

Perley, R. A. (1981b), "VLA hybrid configuration u-v plane coverage", VLA Scientific Memorandum No. 139, NRAO.

Perley, R. A. (1988), "Calibration of P Band data", VLA Scientific Memorandum No. 159, NRAO.

Perley, R. A., Bridle, A. H., and Willis, A. G. (1984), "High-resolution VLA observations of the radio jet in NGC 6251", *Astrophys. J. Suppl. Ser.*, **54**, 291–334.

Perley, R. A. and Erickson, W. C. (1984), "A proposal for a large, low frequency array located at the VLA site", VLA Scientific Memorandum No. 146, NRAO.

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

Readhead, A. C. S., Kemp, M. C., and Hewish, A. (1978), "The spectrum of small-scale density fluctuations in the solar wind", *Mon. Not. Roy. Astron. Soc.*, **185**, 207–225.

Readhead, A. C. S., Walker, R. C., Pearson, T. J., and Cohen, M. H. (1980), "Mapping radio sources with uncalibrated visibility data", *Nature*, **285**, 137–140.

Readhead, A. C. S. and Wilkinson, P. N. (1978), "The mapping of compact radio sources from VLBI data", *Astrophys. J.*, **223**, 25–36.

Rice, S. O. (1944), "Mathematical analysis of random noise", *Bell Sys. Tech. J.*, **23**, 282; (reprinted in *Selected Papers on Noise and Stochastic Processes*, N. Wax, Ed., Dover, New York, 1954).

Rice, S. O. (1945), "Statistical properties of random noise currents", *Bell Sys. Tech. J.*, **24**, 46; (reprinted in *Selected Papers on Noise and Stochastic Processes*, N. Wax, Ed., Dover, New York, 1954).

Rice, S. O. (1948), "Statistical properties of a sine wave plus random noise", *Bell Sys. Tech. J.*, **27**, 109.

Rickett, B. J. (1977), "Interstellar scattering and scintillation of radio waves", *Ann. Rev. Astron. Astrophys.*, **15**, 479–504.

Roberts, D. H., Gabuzda, D. C., and Wardle, J. F. C. (1987), "Linear polarization structure of the BL Lacertae object OJ 287 at milliarcsecond resolution", *Astrophys. J*, **323**, 536–542.

Roberts, D. H., Potash, R. I., Wardle, J. F. C., Rogers, A. E. E., and Burke, B. F. (1984), "Milliarcsecond polarization measurements", in *VLBI and Compact Radio Sources*, R. Fanti, K. Kellermann, and G. Setti, Eds., IAU Symposium No. 110, D. Reidel (Dordrecht, Holland), pp. 35–38.

Roberts, D. H. and Wardle, J. F. C. (1986), "Polarization distributions in compact radio sources", in *Quasars*, G. Swarup and V. K. Kapahi, Eds., IAU Symposium No. 119, D. Reidel (Dordrecht, Holland), pp. 141–147.

Rogers, A. E. E. (1968), "Spectral line interferometry and interferometer noise analysis", Technical Report No. 441, MIT Lincoln Labs.

Rogers, A. E. E. (1976), "Theory of two-element interferometers", in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, pp. 139–157.

Rogstad, D. H. and Shostak, G. S. (1971), "Aperture synthesis study of neutral hydrogen in the galaxy M101", *Astron. Astrophys.*, **13**, 99–115.

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

Rots, A. H. (1988), *A Short Guide for VLA Spectral Line Observers (Edition 7.0)*, NRAO.

Rowson, B. (1963), "High resolution observations with a tracking radio interferometer", *Mon. Not. Roy. Astron. Soc.*, **125**, 177–188.

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

Ruze, J. (1966), "Antenna tolerance theory—A review", *Proc. IEEE*, **54**, 633–640.

Rydbeck, O. E. H. and Hjalmarson, Å. (1985), "Radio observations of interstellar molecules", in *Molecular Astrophysics*, G. H. F. Diercksen, W. F. Huebner and P. W. Langhoff, Eds., NATO ASI Series C Vol. 157, D. Reidel (Dordrecht, Holland), pp. 45–175.

Ryle, M. (1952), "A new radio interferometer and its application to the observation of weak radio stars", *Proc. Roy. Soc.*, **211A**, 351–375.

Ryle, M. (1962), "The new Cambridge radio telescope", *Nature*, **194**, 517–518.

Ryle, M. (1972), "The 5-km radio telescope at Cambridge", *Nature*, **239**, 435–438.

Schwab, F. R. (1978), "Suppression of aliasing by convolutional gridding schemes", VLA Scientific Memorandum No. 129, NRAO.

Schwab, F. R. (1979), "Quantization correction of VLA correlation measurements", VLA Computer Memorandum No. 150, NRAO.

Schwab, F. R. (1980a), "Optimal gridding", VLA Scientific Memorandum No. 132, NRAO.

Schwab, F. R. (1980b), "Adaptive calibration of radio interferometer data", *Proc. S.P.I.E.*, **231**, 18–25.

Schwab, F. R. (1981), "Robust solutions for antenna gains", VLA Scientific Memorandum No. 136, NRAO.

Schwab, F. R. (1984a), "Optimal gridding of visibility data in radio interferometry", in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge University Press (Cambridge, England), pp. 333–346.

Schwab, F. R. (1984b), "Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry", *Astron. J.*, **89**, 1076–1081.

Schwab, F. R. and Cotton, W. D. (1983), "Global fringe search techniques for VLBI", *Astron. J.*, **88**, 688–694.

Schwarz, U. J. (1978), "Mathematical-statistical description of the iterative beam removing technique (method CLEAN)", *Astron. Astrophys.*, **65**, 345–356.

Schwarz, U. J. (1979), "The method 'CLEAN'—use, misuse and variations", in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., D. Reidel (Dordrecht, Holland), pp. 261–275.

Scott, P. F. and Ryle, M. (1972), "A rapid method for measuring the figure of a radio telescope reflector", *Mon. Not. Roy. Astron. Soc.*, **178**, 539–545.

Segalovitz, A. and Frieden, B. R. (1978), "A 'CLEAN'-type deconvolution algorithm", *Astron. Astrophys.*, **70**, 335–343.

Shannon, C. E. (1949), "Communication in the presence of noise", *Proc. IRE*, **37**, 10–21.

Skilling, J. and Bryan, R. K. (1984), "Maximum entropy image reconstruction: general algorithm", *Mon. Not. Roy. Astr. Soc.*, **211**, 111–124.

Sowinski, K. and Braun, R. (1988), "Interim spectral line support", VLA memorandum dated August 12, 1988, NRAO.

Spangler, S. R. (1982), "Correction of VLA K-band amplitudes for atmospheric attenuation", VLA Scientific Memorandum No. 143, NRAO.

Sramek, R. A. (1983), "VLA phase stability at 22 GHz on baselines of 100 m to 3 km", VLA Test Memorandum No. 143, NRAO.

Staelin, D. H. (1974), "The detection and measurement of radio astronomical signals", MIT Dept. Elec. Eng., Lecture Notes.

Steer, D. G., Dewdney, P. E., and Ito, M. R. (1984), "Enhancements to the deconvolution algorithm 'CLEAN'", *Astron. Astrophys.*, **137**, 159–165.

Stratton, J. A. (1935), "Spheroidal functions", *Proc. Nat. Acad. Sci. U.S.A.*, **21**, 51–56.

Švestka, Z. (1976), *Solar Flares*, D. Reidel (Dordrecht, Holland).

Swenson, G. W., Jr. and Mathur, N. C. (1968), "The interferometer in radio astronomy", *Proc. IEEE*, **56**, 2114–2130.

Thompson, A. R. (1973), "The bandwidth effect ('delay beam') for a synthesis array and related requirements for the IF filter characteristics", VLA Electronics Memorandum No. 118, NRAO.

Thompson, A. R. (1980), "Closure accuracy and related instrumental tolerances", VLA Electronics Memorandum No. 192, NRAO.

Thompson, A. R. (1982a), "The response of a radio astronomy synthesis array to interfering signals", *IEEE Trans. Antennas Propagat.*, **AP-30**, 450–456.

Thompson, A. R. (1982b), "The effects of bandwidth and similar parameters", Lecture No. 5 in *Synthesis Mapping*, A. R. Thompson and L. R. D'Addario, Eds., NRAO (Green Bank, WV).

Thompson, A. R. and Bracewell, R. N. (1974), "Interpolation and Fourier transformation of fringe visibilities", *Astron. J.*, **79**, 11–24.

Thompson, A. R., Clark, B. G., Wade, C. M., and Napier, P. J. (1980), "The Very Large Array", *Astrophys. J. Suppl. Ser.*, **44**, 151–167.

Thompson, A. R. and D'Addario, L. R. (1982), "Frequency response of a synthesis array: performance limitations and design tolerances", *Radio Science*, **17**, 357–369.

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

Tiuri, M. E. (1964), "Radio astronomy receivers", *IEEE Trans. Ant. Propagat.*, **AP-12**, 930–938.

Tiuri, M. E. and Räisänen, A. V. (1986), "Radio-telescope receivers", Chapter 7 in J. D. Kraus, *Radio Astronomy*, Second Edition, Cygnus-Quasar Books (Powell, Ohio).

Uson, J. (1986), "Atmospheric opacity at the VLA", VLA Scientific Memorandum No. 157, NRAO.

Van Ardenne, A. (1979), "Loss of sensitivity with Westerbork as a digital tied array, a first study", Netherlands Foundation for Radio Astronomy, Note No. 289.

Van Ardenne, A. (1980), "Calibrating the tied array in the case of analog summation of analog signals", Netherlands Foundation for Radio Astronomy, Note No. 315.

van de Hulst, H. C. (1945), "Radio waves in space", *Ned. Tijdschr. Natuurk.*, **11**, 201.

van der Kruit, P. C. and Shostak, G. S. (1982), "Studies of nearly face-on spiral galaxies. I. The velocity dispersion of the H I gas in NGC 3938", *Astron. Astrophys.*, **105**, 351–358.

Van Vleck, J. H. and Middleton, D. (1966), "The spectrum of clipped noise", *Proc. IEEE*, **54**, 2–19.

Vinokur, M. (1965), "Optimization in the search for a sinusoid of known period in the presence of noise. Application to radio astronomy", *Ann. D'Ap.*, **28**, 412–445.

von Hoerner, S. (1987), "What next? Suggestions for future radio telescopes", NLSRT Memorandum No. 2, NRAO, Charlottesville; also see later memoranda in this series, which relates to the design of large, steerable radio telescopes.

Wade, C. M. (1970), "Precise positions of radio sources. I. Radio measurements", *Astrophys. J.*, **162**, 381–390.

Walker, R. C. (1981), "The multiple-point fringe-rate method of mapping spectral-line VLBI sources with application to H_2O masers in W3–IRS5 and W3(OH)", *Astron. J.*, **86**, 1323–1331.

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

Walker, R. C., Matsakis, D. N., and Garcia-Barreto, J. A. (1982), "H_2O masers in W49N. I. Maps", *Astrophys. J.*, **255**, 128–142.

Weast, R. C. and Selby, S. M. (1975), *Handbook of Tables for Mathematics*, Revised Fourth Edition, CRC Press (Boca Raton, FL), p. 865.

Webster's Third New International Dictionary of the English Language (1971), G. & C. Merriam Company (Springfield, MA).

Wernecke, S. J. and D'Addario, L. R. (1976), "Maximum entropy image reconstruction", *IEEE Trans. Computers*, **C-26**, 351–364.

Wolszczan, A. and Cordes J. M. (1987), "Interstellar interferometry of the pulsar PSR 1237+25" *Astrophys. J.*, **320**, L35–L39.

Woolf, N. J. (1982), "High resolution imaging from the ground", *Ann. Rev. Astron. Astrophys.*, **20**, 367–398.

Yen, J. L. (1985), "Image reconstruction in synthesis radio telescope arrays", in *Array Signal Processing*, S. Haykin, Ed., Prentice–Hall (Englewood Cliffs, NJ), pp. 293–350.